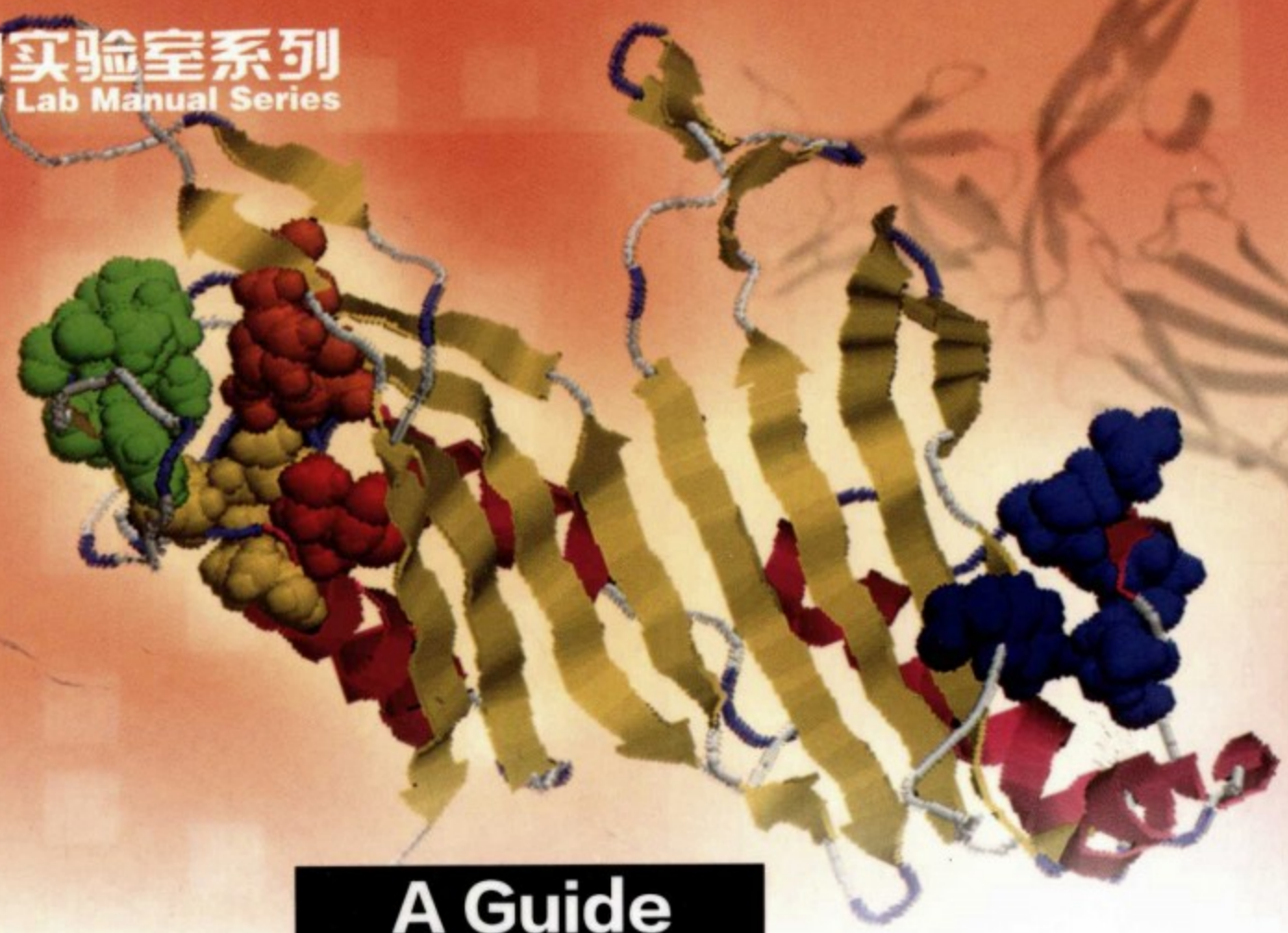
生物实验室系列
Biology Lab Manual Series



A Guide

to Protein Structure Prediction

蛋白质结构预测

实验指南

岳俊杰　冯 华　梁 龙　主编

0.1-33
8

化学工业出版社

内容导读

● 蛋白质结构预测是一种不依赖晶体培养、迅速、简便的获得蛋白质结构的方法，对于分子生物学、蛋白质工程、药物设计等领域研究工作具有重要的意义。

● 本书涵盖了蛋白质结构预测的主流手段与方法，包括同源建模、折叠模式识别、分子对接、蛋白质-蛋白质复合物的结构模拟、分子结构模拟等，以蛋白质结构研究的具体操作和经典实例为素材，使得蛋白质结构预测不仅仅停留在理论阶段，让读者能够轻松地掌握蛋白质结构预测的应用技术。

本书可供从事蛋白质科学研究的研究生、高年级本科生及其他研究人员参考阅读。

ISBN 978-7-122-07755-4

销售分类建议：生物 / 生物技术

定价：38.00 元

Q510.1-33
Y948

生物实验室系列

蛋白质结构预测实验指南

岳俊杰　冯　华　梁　龙 主编

Q510.1-33
Y948

化学工业出版社
·北京·

图书在版编目(CIP)数据

蛋白质结构预测实验指南 / 岳俊杰，冯华，梁龙主编.
—北京：化学工业出版社，2010.4
（生物实验室系列）
ISBN 978-7-122-07755-4

Ⅰ. 蛋… Ⅱ. ①岳… ②冯… ③梁… Ⅲ. 蛋白质–生物结构–实验–指南 Ⅳ. Q510.1–33

中国版本图书馆 CIP 数据核字（2010）第 024203 号

责任编辑：傅四周　　文字编辑：周　倜
责任校对：陈　静　　装帧设计：关　飞

出版发行：化学工业出版社（北京市东城区青年湖南街 13 号　邮政编码 100011）
印　　装：大厂聚鑫印刷有限责任公司
720mm×1000mm　1/16　印张 8½　字数 188 千字　　2010 年 4 月北京第 1 版第 1 次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899
网　　址：http: //www.cip.com.cn
凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：38.00 元

版权所有　违者必究

编 者 名 单

主　　编：岳俊杰　冯　华　梁　龙

编　　者：（按姓氏汉语拼音排序）

冯　华　创腾科技有限公司
郝格非　华中师范大学化学学院
李北平　军事医学科学院生物工程研究所
梁　龙　军事医学科学院生物工程研究所
刘　健　苏州大学材料与化学化工学部
刘　明　嘉和生物药业有限公司
吕　炜　创腾科技有限公司
田　然　创腾科技有限公司
徐　涛　创腾科技有限公司
岳俊杰　军事医学科学院生物工程研究所

出版者的话

21 世纪是生命科学的世纪，这已成为人们的共识。

生命科学随着人类对自身和自然的认识、探索而萌芽，随着人类生产和科学实践的进步而发展。现代生命科学包括生物学、医学、农学等传统学科领域以及生物学、生物技术与环境科学乃至社会科学等其他学科相互渗透、交叉而产生的新型学科体系。20 世纪后叶现代生物科学尤其是分子生物学取得了一系列突破性成就，使得生命科学在自然科学体系中的位置发生了革命性的变化，成为 21 世纪的带头学科。人们对生命科学也寄予了无限的期望，希望能够解决人类社会所面临的人口膨胀、资源匮乏、疾病危害、环境污染和生态破坏等一系列重大问题。

回顾生命科学的发展历程，实验技术一直起着非常重要的促进作用。如 17 世纪 Leeuwenhoek 等人发明并应用显微镜技术，直接催生了“细胞学说”的建立和发展；1973 年 Cohn 和 Boyer 完成了 DNA 体外重组实验，标志着基因工程的肇始；1988 年 Kary Mullis 发明的 PCR 技术甚至使生命科学产生了飞跃性的发展。可以说，生命科学无时无刻离不开实验，实验是开启神奇的生命王国大门的钥匙。没有实验技术的不断进步，也就没有生命科学今天的巨大发展；同时，生命科学的发展又对实验技术提出了更高的要求，进一步刺激了后者的不断进步。生命科学正是在“实验催生和验证着基础理论，理论指导和发展了实验技术”的不断循环中从必然王国走向自由王国。

工欲善其事，必先利其器。为了有助于生命科学工作者更多地了解相关实验技术和仪器设备，更好地设计实验方案，更有效地开展实验过程，更合理地处理实验结果，化学工业出版社组织出版了“生物实验室系列”图书。系列图书在整体规划的基础上，本着“经典、前沿、实用，理论与技术并重”的原则组织编写，分批出版。

在题材上，系列图书涵盖综合实验技术和单项实验技术两个方面。其中综合实验技术既有以实验目的为题，如“蛋白质化学分析技术”，内容纵向覆盖多项实验技术；也有以某一生命学科领域的综合实验技术为题，如“发酵工程实验技术”、“生物化学实验技术”等。而单项实验技术则以深入介绍某一专项技术及其应用为主，在阐述其基本原理的基础上，横向介绍该项技术在多个领域的应用，如“双向电泳技术”、“流式细胞术”等。

在内容上，系列图书主要有以下两个显著特点。一是强调先进性——除了系统介绍常用和经典实验技术以外，特别突出了当前该领域实验手段的新理论、新技术、新发展，为国内专业人员起到借鉴和引导作用。二是强调可操作性——对于每一项实验技术，系统介绍其原理方法、设备仪器和实验过程，让读者明了实验的目的、方案设计以及具体步骤和结果处理，以期起到实验指南的作用。

本系列图书坚持质量为先，开拓国内和国际两个出版资源。一方面，约请国内相关领域兼具理论造诣和丰富实验室工作经验的专家学者编著；另一方面，时刻关注国际生命科学前沿领域和先进技术的进展，及时引进（翻译或影印）国外知名出版社的

权威力作。

《生物实验室系列》图书的读者对象设定为国内从事生命科学及生物技术和相关领域（如医学、药学、农学）的专业研究人员，企业或公司的生产、研发、管理技术人员，以及高校相关专业的教师、研究生等。

我们殷切希望《生物实验室系列》图书的出版能够服务于我国生命科学的发展需要，同时热忱欢迎从事和关心生命科学的广大科技人员不仅能对已出版图书提供宝贵意见和建议，也能对系列图书的后续题目设计贡献良策或推荐作者，以便我们能够集思广益，将这一系列图书沿着可持续发展的方向不断丰富品种，推陈出新。

谨向所有关心和热爱生命科学，为生命科学的发展孜孜以求的科学工作者致以崇高的敬意！

祝愿我国的科技事业如生命之树根深叶茂，欣欣向荣！

化学工业出版社

生物·医药出版分社

前　言

蛋白质是一切生物借以表现生命的最重要的基本单元，是自然界存在的小型的自动机器。任何一个生命体的繁衍、新陈代谢、运动等，都需要数十亿个蛋白质分子的协调行动才能得以顺利进行。随着人类功能基因组研究的展开，科学家对于基因的研究焦点，已由基因测序转移到基因表达产物——蛋白质上。为了执行特定的生物功能，每一个蛋白质分子都有一个独特的三维结构。研究蛋白质的结构，有助于了解蛋白质的功能，了解蛋白质的作用机制以及了解蛋白质与其他分子之间的相互作用。对于全新或功能未知的蛋白质分子，通过结构分析，可以进行功能诠释，指导设计生物学实验来进行功能研究。通过分析蛋白质的结构，确认功能位点，为设计新的蛋白质或改造蛋白质提供可靠的依据。到目前为止，确定蛋白质三维结构的方法主要分为两大类，其一是利用实验的方法来测定；其二则是利用计算机技术，根据现有理论和已知的序列等信息进行蛋白质的结构预测。测定蛋白质三维结构的方法分为三大类：（1）X 射线晶体衍射图谱法，（2）核磁共振方法，（3）电镜方法。

近年来分子生物学的快速发展使蛋白质序列数据库的数据量得到快速增长。而利用实验方法测定蛋白质三维结构则由于耗时长、成本高以及方法在技术上的困难，而使解析的蛋白质结构比已知的蛋白质序列要少得多。另一方面，随着 DNA 测序技术的发展，人类基因组及更多的模式生物基因组将被完全测序，DNA 序列数量将会急增，由于 DNA 序列分析技术和基因识别方法的进步，从 DNA 可以很方便地推导出大量的蛋白质序列。这意味着已知序列的蛋白质数量和已测定结构的蛋白质数量的差距将会越来越大。因此通过实验方法获得蛋白质三维结构越来越无法满足人们对蛋白质结构和功能关系研究的要求，迫切需要一种不依赖实验而又有一定准确性的获取蛋白质三维结构信息的手段，蛋白质结构预测在后基因组时代就显得格外重要。

目前国内外关于蛋白质结构预测的专著已有不少，但是大都偏重于理论方面。对于从事生命科学与生物技术研究的广大科研工作者来说，需要的是实用性强、具有可操作性的技术类著作，而这方面的资料几乎是空白。为了使在生物医学领域的科技人员及研究生能够循序渐进地掌握蛋白质结构预测的方法，我们组织了一批在一线从事相关工作的年轻博士出版了这本小册子。主要向广大读者详细介绍蛋白质结构模建、蛋白质的分子对接、蛋白质-蛋白质相互作用的模拟、蛋白质结构的分子动力学模拟等操作。每一种分析技术都选择两种软件，一是在业界声誉最好的商用软件，二是应用最普遍的免费软件。计算分析的对象都是各位编者个人在实践工作中的具体的分析对象。期望通过阅读本手册，可以使从事生命科学与生物技术研究的广大科研工作者能够掌握相关的技术要领，学会掌握这些方法，为他们的科研工作服务。

本书的第一章由岳俊杰撰写，第二章、第四章由岳俊杰、梁龙、李北平撰写，第三章由岳俊杰、冯华、徐涛撰写，第五章由冯华、刘明、田然、吕炜撰写，第六章由冯华、刘明、徐涛、吕炜撰写，第七章由冯华、田然、刘明、郝格非、刘健撰写。

编写过程得到了军事医学科学院生物工程研究所和创腾科技有限公司的大力协助，在此表示感谢。

限于编者的水平和精力，再加上时间仓促，书中可能存在许多不妥之处，敬请广大读者批评和指正。热情期盼把您的宝贵意见函至 yue_junjie@126.com。

编　者

2010 年 3 月

目　录

第一章　蛋白质结构概述……1
　第一节　蛋白质的结构层次……1
　　一、一级结构……1
　　二、二级结构……1
　　三、三级结构（折叠模式）……2
　　四、四级结构……2
　第二节　蛋白质结构的测定方法……2
　　一、X 射线晶体学……2
　　二、核磁共振……4
　　三、结构分析的其他方法……4
　第三节　蛋白质结构预测的意义……5
第二章　蛋白质二级结构预测……7
　第一节　蛋白质二级结构预测概述……7
　第二节　蛋白质二级结构预测实例……8
第三章　利用同源模建法预测蛋白质结构……10
　第一节　同源模建的基本原理及基础……10
　第二节　同源模建预测蛋白质结构的基本步骤……11
　　一、参考蛋白搜索……11
　　二、确定结构保守区……11
　　三、蛋白主链的模建……11
　　四、侧链安装……12
　　五、优化处理……12
　　六、合理性检测……12
　第三节　利用 SWISS-MODEL 服务器进行蛋白质结构同源模建操作实例……12
　　一、登录服务器，输入目标序列……12
　　二、模建结果与分析……13
　第四节　利用 Discovery Studio 结构模拟系统进行同源模建操作实例……15
　　一、识别模板，比对模板……16
　　二、将模板进行比对……19
　　三、将目标序列与模板比对……23
　　四、使用 MODELER 构建目标序列的 3D 模型……25
　　五、模型评估……26
第四章　利用折叠模式识别法预测蛋白质结构……32
　第一节　折叠模式识别法的基本原理……32

第二节　折叠模式识别法预测蛋白质结构的操作实例 33
一、程序的安装 33
二、Threader 程序的运行 33
三、结果分析 34
四、过滤假阳性结果 36
第五章　蛋白质分子对接技术 37
第一节　蛋白质分子对接概述 37
第二节　利用 LibDock 进行分子对接的基本操作方法及实例 37
一、准备分子对接体系，执行分子对接计算 38
二、分析对接结果 43
第三节　利用 Flexible Docking 进行受体和配体柔性的分子对接实例 46
一、准备对接的分子体系，完成分子对接计算 46
二、分析配体对接结果 49
第四节　利用 AutoDock 程序进行分子对接的操作及研究实例 50
一、AutoDock 的安装 50
二、AutoDock 的基本应用 52
三、实例分析 68
第六章　蛋白质-蛋白质复合物的结构模拟 70
第一节　蛋白质与蛋白质对接的意义 70
第二节　利用 ZDOCK 来研究蛋白质-蛋白质的对接 70
一、设定一次 ZDOCK 计算 71
二、分析 ZDOCK 结果 72
三、利用 RDOCK 优化对接构型 75
四、使用 RMSD 分析结合界面的氨基酸 76
第三节　利用 Hex 研究大分子与大分子之间的相互作用 78
一、Hex 的安装 79
二、Hex 的基本应用 79
三、对接实例 84
第七章　蛋白质结构的分子动力学模拟 85
第一节　蛋白质分子动力学模拟概述 85
第二节　Discovery Studio 平台中分子动力学操作实例 86
一、准备分子动力学体系，执行分子动力学计算 86
二、分析分子动力学计算结果 90
第三节　利用 NAMD 进行分子动力学计算的操作及研究实例 95
一、NAMD 简介 95
二、NAMD 程序的安装与使用 95
三、构建模型 97
四、分子动力学模拟 102
五、可视化与数据处理 108
第四节　利用 GROMACS 进行分子动力学模拟操作 113

一、GROMACS 介绍……113
二、GROMACS 的下载及安装……113
三、GROMACS 的基本操作……114
本书主要参考文献…… 123

第一章

蛋白质结构概述

蛋白质是生命活动的主要承担者，一切生命活动无不与蛋白质密切相关。蛋白质所具有的生物学功能在很大程度上取决于蛋白质分子的结构。蛋白质是由一条或几条多肽组成的大分子，每条多肽链都是一个线性的氨基酸链。蛋白质是生物大分子，具有明显的结构层次性，由低层到高层可分为一级结构、二级结构、三级结构和四级结构。一级结构也称为初级结构，二级结构、三级结构和四级结构被称为高级结构。

第一节　蛋白质的结构层次

一、一级结构

蛋白质的一级结构是指肽链的氨基酸组成及其排列顺序，包括二硫键的连接关系。氨基酸序列是蛋白质分子结构的基础，它决定蛋白质的高级结构。氨基酸通过肽键连接在一起，肽键一般采取反式构型，也就是说羰基氧原子和临近氨基酸的氨基氢原子彼此指向相反的方向。肽键本身是刚性的，但是其他键则具有很大的柔性，从而使多肽主链能够在空间折叠。尽管反式构型依然是优先的，但脯氨酸的残余基团结合到多肽链的主链上，因而这样的残基可以形成顺式的肽键，使羰基氧原子和邻近氨基酸的氨基氢原子指向同一个方向。这对肽链主链的折叠有很重要的影响，而且用其他氨基酸残基取代脯氨酸后不可避免地会对整体结构有重大的影响。正是因为这个原因，脯氨酸通常在蛋白质结构中是高度保守的。与此类似的是，甘氨酸也是很重要的残基，因为它具有很小的残余基团（只有一个氢原子），这样使它比其他氨基酸残基具有更大柔性度。半胱氨酸残基也是高度保守的，因为它有形成二硫键的能力，可以将分离的多肽链连接在一起。

二、二级结构

蛋白质的二级结构是由分子内氢键产生的有规则的和重复的局部构型。二级结构有时候涉及极性侧链（如丝氨酸和苏氨酸残基的侧链），但是多肽主链本身也是极性的，因为NH基团可以作为氢原子供体，C═O基团可以作为氢原子受体。肽键贯穿多肽主链的有规则分布有利于重复有序的结构形成。α螺旋和β折叠片是两种最普通的结构。α螺旋通常是右手螺旋，当两个相距四个残基的肽单元之间形成氢键时就会出现α螺旋。α螺旋使肽键排成了一行。α螺旋的大小变化是4～40个残基，对应于1～11圈螺旋。相反，β折叠片是在多肽链的键角完全扩展的区域（这些区域称为β折叠股）形成的。几个β折叠股可以排列成平行（parallel）、反平行（antiparallel）或者混合型的阵列，相邻折叠股之间的肽单元形成了氢键。α螺旋和β折叠片可以通过一些连接区域结合在一起，这些连接区域采取的是它本身的二级结构，这种二级结构称为回折（turn）结构。如果连接区域没有出现氢键，这样的连接区域成为连接回环（loop）。一个蛋白质的内核通常是富含二级结构的，而环区一

般出现在表面。

三、三级结构（折叠模式）

一个多肽的三级结构，或者折叠模式是它的总体外形，反映了二级结构和模体堆积在一起形成致密结构域的方式。一个结构域可以被看成是一个多肽链中的一部分，这部分可以独立折叠成一个稳定三维结构。结构域也可以被定义为功能的单元。一个蛋白质可以包含一个结构域，也可以包含几个结构域。在后面的这一种情况下，不同的结构域可以按照与蛋白质的整体生物学功能有关的前后联系来执行完全不同的生物学功能。

四、四级结构

很多蛋白质只有一条多肽链，但是其他的很多蛋白质则包含多条多肽亚基。这些亚基的装配方式决定了蛋白质的四级结构。一个多结构域蛋白质和具有几个不同多肽亚基的蛋白质之间没有功能上的差别，很多蛋白质以这两种形式存在。例如，大多数转录因子都是一个具有 DNA 结合和转录激活结构域的单条多肽，另外一个则是由不同的亚基装配成的。实际上，相互作用的亚基形成的转录因子的装配正是酵母双杂交系统检测二元蛋白质作用的基础。蛋白质亚基可以通过非共价键相互作用，也可以通过多肽链间二硫键结合在一起。

第二节　蛋白质结构的测定方法

在没有任何形式的结构数据来源的情况下，从一个蛋白质的序列开始，从头预测这个蛋白质的三级结构在目前依然是不可能的。因而，在所有可信程度的情况下都可以有把握地确定一个在某种程度上特征完全未知的蛋白质的结构的唯一方法就是通过实验手段来测定。有两种主要的技术可以用于这个目的，这些技术就是 X 射线晶体学（X-ray crystallography，XRC）和核磁共振（nuclear magnetic resonance，NMR）谱。蛋白质结构库（PDB）中超过 98%的结构是采用上述两种方法中的一种测定的，余下的 2%中的绝大多数是基于 X 射线晶体学结构和 NMR 结构的理论模型。全部结构中有不到 100 个是利用其他方法，如中子衍射、电子衍射和电子显微镜得到的。

X 射线晶体学和 NMR 两种方法都有非常苛刻的要求，对于每一个蛋白质来说，必须根据经验来确定准确的实验条件。X 射线晶体学中蛋白质晶体的制备被认为是一门“艺术”，测定蛋白质结构的很多尝试都是因为拿不到合适的晶体而失败。对于 NMR 谱，则是蛋白质以溶液状态测定结构，蛋白质必须在高浓度下是可溶而且稳定的，而且在这样的条件下不能聚集或变性。这两种技术都需要对大量的数据进行收集和处理，然后是艰苦的模型装配以便于产生原子坐标，装配的模型要与实验得出的结果相一致。

一、X 射线晶体学

X 射线晶体学利用的是这样一个原理，即当 X 射线经过蛋白质晶体时，它会以一种可预测的方式被散射或衍射。X 射线在遇到电子时会发生衍射，因此散射的特征依赖于出现在每个原子中电子的数量和原子在空间的排列。与其他的波一样，衍射的 X 射线彼此之间会发生正向的或负向的干涉。当蛋白质分子有规则地排列在一个晶体中时，由不同分子的等价原子散射产生的在同一方向的 X 射线会发生相互作用，这将在探测仪上生成一个斑点图案，这种斑点图案称为反射。这些衍射图案可以用于构建分子的电子云的三维图像，这

种电子云的三维图像被称为电子密度图。蛋白质的结构模型就是搭建在这样的电子密度图中。

精确的结构测定需要一个高度有序的晶体，这样的晶体能够使 X 射线产生强烈衍射。蛋白质晶体生长特别困难，这是结构测定中一个不可忽视的瓶颈。疏水蛋白质或含有疏水结构域的蛋白质是最难结晶的，这也就是 PDB 数据库中只有很少几个完整膜蛋白结构的原因，因为这些蛋白质都有疏水的跨膜结构域。自动化结晶工作台的开发有助于扩大 X 射线晶体学通量，这样就可以在同一时间尝试试验数千个不同的参数条件，比如不同的蛋白质浓度、盐浓度、温度和 pH 等，以确定最佳结晶条件。少量的样品也可以采用，这样可以进行那些含量不丰富的蛋白质的结晶研究。

获得蛋白质晶体之后，X 射线晶体学面对的下一个问题是从衍射图案计算一个电子密度图。这需要三方面的信息：入射 X 射线的波长、散射的 X 射线的振幅（可以通过反射的强度测定出来）和衍射的相位。不幸的是，相位不能从反射的图案中测定出来，这就带来了相位问题。有时候可以从已经解析的、存放在蛋白质数据库中的相关结构中来“借”相位，这种处理手段称为分子置换法。不过，多数时候需要进行进一步的实验来测定衍射相位。标准的处理过程是制备包含重金属原子的同形晶体，也就是在相同总体结构的晶体中结合上较重的原子，这样可以产生另外一种不同的衍射图案。包含重金属原子的同形晶体可以通过在重金属的盐溶液中浸泡来制备，经过浸泡，重金属原子能够扩散到原先由溶剂分子占据的空间，并与蛋白质分子中某些确定的位置结合。重金属原子比那些蛋白质分子中正常出现的原子更强烈地使 X 射线发生衍射。通过比较多种不同的同型晶体产生的反射（称为多对同晶置换法，multiple isomorphous replacement，MIR），就可以确定这些重原子的位置，这样就可以推导出未置换晶体中衍射的相位。每一个反射的全部描述即波长、振幅、相位，它们就是结构因子。

测定结构因子的相位也可以用异常散射法来实现。当蛋白质分子中的重金属原子在遇到波长接近其天然吸收边限的 X 射线照射时，就会导致这些重金属原子会以附加 X 射线的形式重新发射出其中的部分能量，这就是发生了异常散射。异常散射的幅度随着入射 X 射线的波长而变化，因此一种含有重金属原子的晶体可以被几种不同波长的 X 射线照射，而产生几种不同的衍射图案，从这些衍射图案中，散射的相位就可以计算出来。这就是单对同晶置换加异常散射（single isomorphous replacement with anomalous scattering，SIRAS）和多波长异常散射（multiple wavelength anomalous dispersion，MAD）等技术的基础。一种改进的处理手段可使蛋白质晶体无须浸泡，这种方法是将蛋白质在细菌中表达时加入一些重金属原子取代的氨基酸类似物。

最后，需要在电子密度图中搭建一个结构模型。这需要一些更有决定性作用的信息——氨基酸序列，因为通过 X 射线照射不可能明确地区分出碳、氧和氮等原子，所以识别氨基酸侧链很困难。结果得到的模型是一套原子 *XYZ* 坐标，赋予了除氢以外的所有原子。电子密度的数据越多，原子位置的确定程度越高，模型的分辨率也越高。虽然如此，仍然可能会有蛋白质分子中某些区域的原子位置不能精确测定。每个原子都被赋予一个所谓的温度因子，这是确定性的一个度量。温度因子越高，则确定性越低。高温度因子说明了两种情况，一种是无序的一个度量；另外一种是动力学状态。

二、核磁共振

核磁共振（nuclear magnetic resonance，NMR）是因为某些原子核具有磁的性质而发生的一种现象。在 NMR 谱中，这些性质可以用来获得化学的信息。可以认为亚原子的微粒都在绕着它们的轴旋转，并且在很多原子中，这些旋转彼此平衡抵消，所以原子核本身没有总旋转。在氢原子（^{1}H）和某些天然存在的碳、氮的同位素（^{13}C、^{15}N）中，这些旋转不能抵消，这类原子核会拥有一个所谓的磁矩。这样的原子核可以采取两种可能的取向中的一种，在正常的情况下，这两种取向都有相同的能量。但是在一个外加磁场的情况下，这些能级就会分裂，这是因为原子核磁矩的一个取向是与外加磁场平行的，而另外一个取向则不是。如果这种能量间隔存在，而且原子核暴露在某一特定频率的电磁辐射中时，那么这个原子核就会被诱导发生跃迁，从低能量的磁旋转状态跃迁到不优先选择的高能量旋转状态。因为电磁辐射的频率与原子核旋转的频率是一致的，所以这种吸收称为共振。当原子核回到它们的初始定向时，这些原子核会发射出电磁波，这些电磁波可以被测量出来。质子（^{1}H）能够给出最强的信号，这一点正是利用 NMR 谱进行蛋白质结构分析的基础。

NMR 谱可以用于测定溶液中蛋白质的结构，但需要蛋白质是非常可溶而且高度稳定，需要溶液浓度接近 1mmol/L，并且需要的体积比较大（1～5mL）。因为所有成键原子附近的电子都会产生自己的磁场，所以成键原子附近的电子会影响到每个原子核的磁共振频率，因此 NMR 谱可以用来测定结构。外加的磁场的强度必须提高到能够克服电子产生的磁场的对抗或屏蔽，扰动的幅度或化学位移取决于每一个原子核的化学环境。通过这种方法，可辨别不同的氢原子，如甲基和芳香基团的氢原子的辨别。

一维核磁共振实验可以检测到化学位移以及其他的类似于自旋-自旋耦合（spin-spin coupling）这样的屏蔽效应，但是在通常情况下，这些对于描述像蛋白质这样复杂的分子是不够的。可以采用一些由不同时间间隔区分的系列脉冲代替使用单个脉冲。这样可以得到一个二维的 NMR 谱，这种谱携带了额外的峰，这种峰可以显示具有相互作用的原子核对。

当把相互作用的原子核对的各种效应都考虑进去之后，NMR 分析的结果就是一些距离制约的组合，通过这些距离约束，可以用来推算某些特定的原子对之间的距离（包括成键的和没有成键的）。如果足够的距离约束计算出来，那么符合这些数据的蛋白质的结构数量就会变得非常有限。NMR 分析产生的是 10～50 个模型的组合体，而不是一个单一的结构。高精度的 NMR 谱依赖于蛋白质分子在溶剂中的快速翻转，这限制了可以分析的蛋白质的大小，目前可以分析的蛋白质都少于 300 个氨基酸残基。如果质子的 NMR 谱过于密集，分析也可以扩展到其他的原子核（如 ^{13}C 及 ^{15}N）以产生多维和异核的 NMR 谱，这样可以降低数据密度。

三、结构分析的其他方法

除 X 射线晶体学和 NMR 外，有些其他方法也可以为解析的结构提供补充信息，或者分析那些在 X 射线晶体学和 NMR 谱两种情况都不能分析的蛋白质结构。例如，圆二色谱（circular dichroism spectrophotometry，CDS）就是一个很有用的测定蛋白质二级结构的方法。圆二色是一种光学现象。不对称的分子如蛋白质，在左旋圆偏振化光和在右旋圆偏振化光中有不同的吸收谱。圆二色采用的是波长 160～240nm 的光，这个波段的光能够分别对含α螺旋和β折叠的蛋白质产生明显不同和具有特征的谱。圆二色谱不能测定蛋白质的三

级结构，但是对于结构生物学中的 X 射线晶体学和 NMR 谱是个有益的补充。

其他的可以用于研究蛋白质三级结构的方法包括中子衍射和电子衍射。中子衍射比 X 射线衍射（XRD）使用的机会少得多，这是因为中子源不能广泛地得到，而且中子束的流量约比 X 射线束的流量低 10 个数量级。电子衍射用来研究那些能够结晶形成或天然装配成二维排列但不形成有序的三维晶体的蛋白质的结构。电子显微镜的优势就是可以用分析结晶排列的蛋白质相同的方式分析那些单个分子，在不用结晶的情况下测定出大蛋白复合物的结构。电子显微镜的分辨率常常没有 X 射线衍射的分辨率高，但是由于不需要结晶，在只有少量而且不太纯的样品的情况下也可以进行测定。

第三节　蛋白质结构预测的意义

尽管解析蛋白质结构的技术已经取得了非常重大的进展，获得了大量的蛋白质结构，但解析蛋白质结构依然是一个劳动密集型和花费昂贵的过程，而且目前各种实验方法都存在缺陷或限制。X 射线单晶衍射最大的缺点是要测定的蛋白质必须能够形成晶体，而许多蛋白质不能结晶，这就限制了测定的范围。核磁共振方法虽然不需培养蛋白质晶体，但需要蛋白质能够在较高浓度下可溶、稳定、不聚集或变性，更重要的是，核磁共振技术对于较大的蛋白质分子的结构还无能为力。电镜三维重构需要培养二维晶体，而且目前测定的结构的分辨率还不够高，还没有成为测定蛋白质结构的常规方法。

蛋白质空间结构的测定速度还远远赶不上蛋白质序列增长的速度，并且随着基因组测序的进行，二者差距有越来越大的趋势。截止 2009 年 10 月，已知序列的蛋白质已超过 100 万，而已知结构的蛋白质仅有 6 万多。因此蛋白质结构预测成了目前了解蛋白质结构信息的一个非常有效的实际手段。随着基因组和蛋白质组计划的研究进展，大量的具有特殊功能的蛋白质将被发现，对蛋白质空间结构预测的要求也越来越迫切。

除了获得蛋白质结构信息之外，蛋白质结构预测还有重大的理论意义。生命遗传信息存储传递及表达的认识是 20 世纪生物学所取得的最重要的突破，其中的关键问题是由 3 个相连的核苷酸顺序决定蛋白质分子肽链中的 1 个氨基酸，即“三联遗传密码”（“第一遗传密码”）的破译。蛋白质必须有特定的三维空间结构，才能表现其特定的生物学功能。20 世纪 60 年代 Anfinsen 就提出蛋白质分子的一级序列决定其空间结构的论断，这一论断得到多次实验证实并被人们广泛接收，大量实验充分说明蛋白质的氨基酸顺序与其空间结构之间存在着确定的关系。国际上将蛋白质的氨基酸序列与其空间结构的对应关系称之为“第二遗传密码”。但是直到现在，“第二遗传密码”依然没有得到破译，这个问题是蛋白质研究最后几个尚未揭示的奥秘之一。蛋白质结构预测实际上是去从理论上最直接地解决蛋白质的折叠问题，即破译“第二遗传密码”。

近年来随着结构生物学的发展，解析的蛋白质结构越来越多，这就为研究和总结蛋白质结构的规律打下了很好的基础，也为蛋白质的结构预测提供了参考。另外计算机科学与技术的快速进步也大大地促进了蛋白质结构预测的发展。蛋白质结构预测问题虽然还没有最终解决，但是已经取得了一些令人欣喜的成就。目前，蛋白质空间结构预测的方法有 3 种，即从头预测法、比较建模法和折叠识别法。

1．从头预测法

从头预测（ab initio prediction）也称为理论计算预测，是指从蛋白质的一级结构出发，

根据物理化学、量子化学、量子物理的基本原理，利用各种理论方法计算出蛋白质肽链所有可能的构象的能量，然后从中找到能量最低的构象，就是蛋白质的天然构象。这种方法不需要已知结构信息，能够产生全新结构。但是由于计算的难度，这种方法只能用于计算很小的分子或蛋白质分子的局部结构，目前还不能作为一种常用的预测蛋白质结构的方法。现在从头预测主要用作其他预测方法的补充或作为一种优化结构的手段。

2．比较建模法

比较建模法（comparative modeling）也称为同源建模法（homology modeling），是基于知识的蛋白质结构预测方法。根据对 PDB 中的蛋白质进行结构比较的定量研究得知，任何一对蛋白质，只要它们序列的长度达到一定程度，序列相似性超过 30% ，则可以保证它们具有相似的三维结构。因此，对于一个未知结构的蛋白质，如果找到一个已知结构的同源蛋白质，就可以该蛋白质的结构为模板，为未知结构的蛋白质建立结构模型。同源模建通常包括下列主要步骤：模板搜寻、序列比对、结构保守区寻找、目标模型搭建、结构优化和评估等。在目前的三种预测蛋白质结构的方法中，比较建模法是最简单、最为成熟的。很多实验室都可以开展同源模建预测蛋白质结构的工作。

3．折叠识别法

折叠识别法（fold recognition）也称为反向折叠法（inverse fold method）、threading 等。该方法基于这样一个事实，即很多序列没有相关性的蛋白质具有相似的折叠模式。因此可以基于序列结构比对（sequence structure alignment）的预测方法，通过目标蛋白质的氨基酸序列和已知折叠模式的逐一比对，根据特定的计分函数（scoring function）找出最适合目标序列的折叠模式。折叠识别法可以弥补同源模建方法只能依赖序列相似性寻找模板的不足，是目前三种预测蛋白质结构的方法中发展最快也是最有前途的方法。

第二章

蛋白质二级结构预测

第一节　蛋白质二级结构预测概述

蛋白质的二级结构是指多肽链中沿一个方向排列成具有周期性结构的构象，它是蛋白质多肽主链中有规则的重复构象。蛋白质二级结构是蛋白质分子中重要的组成“部件”，是研究蛋白质氨基酸序列和三级结构之间的桥梁。蛋白质二级结构的预测对于蛋白质三维结构的预测具有十分重要的意义。

二级结构预测也常称为三态预测，因为序列中的每一个氨基酸残基都可以归结为螺旋（helix，H）、拉长的折叠股（extended β-strand，E）或卷曲（coil，C）三种状态。早在 20 世纪 60 年代中后期，科学家们就开始了蛋白质二级结构预测方面的研究，到目前已发展了几十种预测方法。这些方法大致上可分为三大类：第一类是统计学方法，第二类是基于立体化学原则的物理化学方法，第三类是神经网络和人工智能方法。

Chou-Fasman 法是统计学方法的典型代表。Chou-Fasman 对已知结构的蛋白质进行了统计处理，计算出 20 种氨基酸出现在 α 螺旋、折叠和卷曲三种构象中的分布情况，然后得到每一种氨基酸在这些二级结构构象中的构象参数。Chou-Fasman 法中的构象参数主要有：氨基酸残基形成 α 螺旋的倾向性因子 Pα；氨基酸残基形成折叠股的倾向性因子 Pβ；氨基酸残基形成卷曲结构的倾向性因子 Pt。氨基酸的构象参数值反映了氨基酸出现某种二级结构中可能性的大小。根据统计的规律，Chou-Fasman 提出了一些二级结构成核、延伸和终止的规则，根据这些规则可以预测已知序列的多肽链的二级结构。

利用物理化学方法预测蛋白质二级结构最著名的是 Lim 方法。该方法在预测蛋白质二级结构时考虑了蛋白质折叠构象的立体化学特征和物理化学性质，比如残基侧链基团的体积大小、亲/疏水性质和所带电荷等因素，另外这一方法还充分考虑了邻近氨基酸残基之间的相互作用情况。此外，根据已经测定的蛋白质结构情况，该方法还总结了形成 α 螺旋或 β 折叠构象的结构模式和立体化学特征。比如，可以看到 α 螺旋结构中经常是一个侧面对着亲水表面，另一侧对着疏水核心。Lim 等人对 α 螺旋和 β 折叠总结了 20 多种亲/疏水分布模式。根据序列中有规律的亲/疏水性残基的分布，可以对蛋白质的二级结构进行预测。在有些情况下该方法的预测准确率相当高（尤其是对较小的蛋白）。

神经网络和人工智能方法是新兴的预测蛋白质二级结构的方法，目前二级结构预测应用最广的神经网络算法是 BP 网络（back-propagation network），即反馈式神经网络算法。它通常是由三层相同的神经元构成的层状网络，使用反馈式学习规则，底层为输入层，中间为隐含层，顶层是输出层，信号在相邻各层间逐层传递，不相邻的各层间无联系，在学习过程中根据输入的一级结构和二级结构的关系的信息不断调整各单元之间的权重，最终目标是找到一种好的输入与输出的映像，并对未知二级结构的蛋白进行预测。这些方法与

传统的方法相比最大的特点是预测二级结构时不仅仅是基于单序列和单个氨基酸，而是利用了多序列。通过从多重序列的比对中学习规则，利用神经网络的方法进行序列的预测。这类算法发展较快，已经有很多成功的应用。

第二节　蛋白质二级结构预测实例

PSIPRED 方法是由英国 David T. Jones 实验室开发的基于神经网络算法的蛋白质二级结构预测软件，在预测过程中利用了多重序列比对并且结合了从数据库中提取的结构和进化信息。它先使用 PSI-BLAST 在数据库中搜索序列的相似蛋白，构建多序列比对，然后在此计算结果基础上进行结构预测。PSIPRED 用两个前向神经网络对来自 PSI-BLAST 的特征图进行分析。序列可以通过互联网用简单的单字母格式或是 FASTA 格式提交，PSIPRED 的预测结果通过电子邮件以文本文件形式发送回来。另外，在电子邮件中会给出一个网址，到那里可以看到被预测蛋白质的图像表示，可视化是用 JAVA 应用程序 PSIPREDview 实现的。PSIPRED 的平均预测准确率为 78%，在多次二级结构预测评估中取得领先，被认为是一种较为准确的蛋白质二级结构预测方法。本节以幽门螺杆菌脂蛋白 lpp20 为例，利用 PSIPRED 程序预测了其二级结构。

第 1 步，启动网络服务器，登录页面 http://bioinf4.cs.ucl.ac.uk:3000/psipred，就会看到如图 2-1 页面。

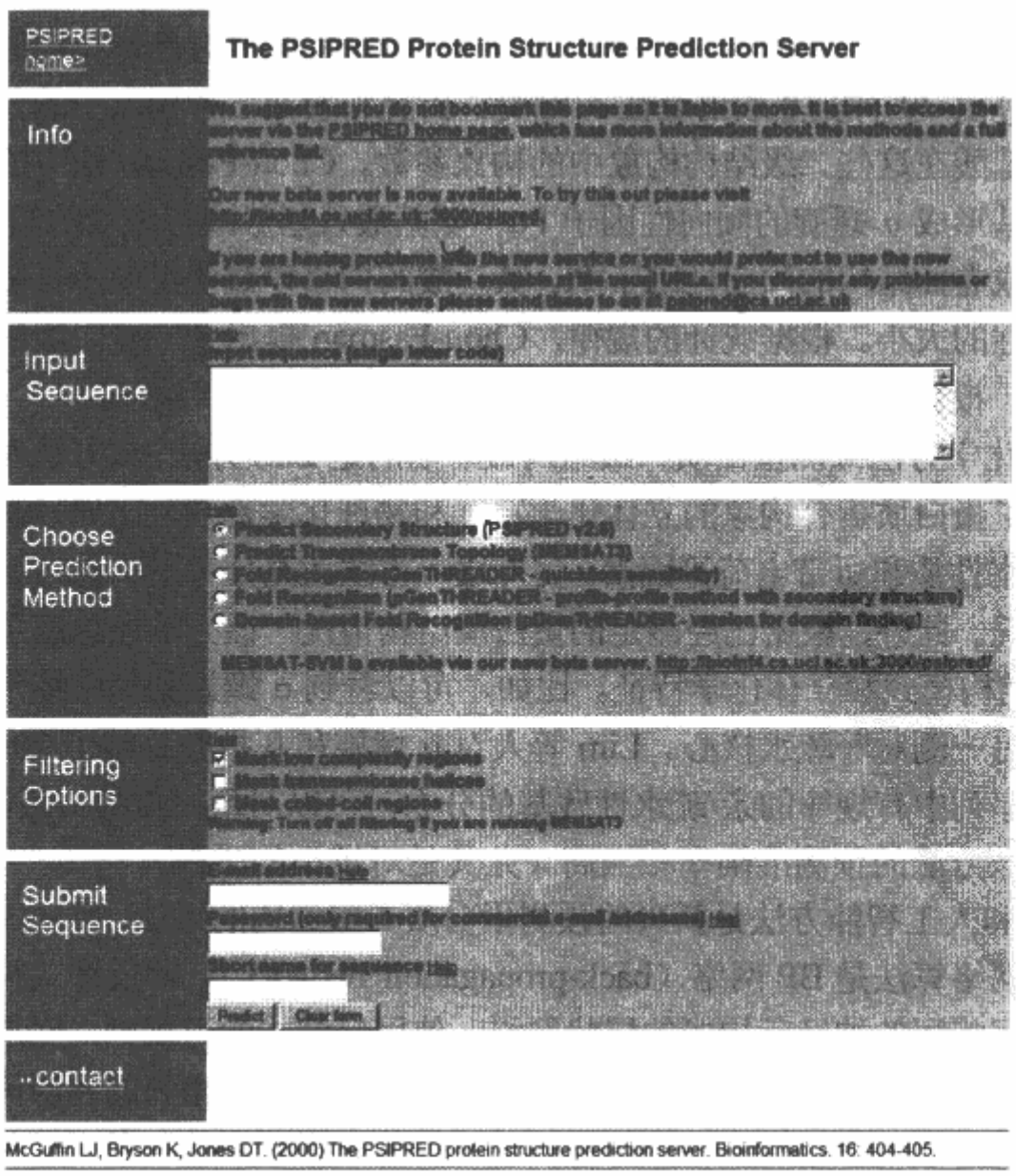


图 2-1　PSIPRED 蛋白质二级结构在线预测窗口

第 2 步，在页面中填入 lpp20 蛋白的序列、自己的电子信箱，并点击 Predict 命令，即可完成预测任务的输送，程序会自己进行计算。

第 3 步，几分钟后，在自己的信箱中会收到预测结果的链接。自己的预测页面也会切换到预测结果。

lpp20 蛋白的二级结构预测结果如图 2-2 所示。结果中不仅包括预测的各个残基所处的构象，而且还给出了预测的可信度评估。

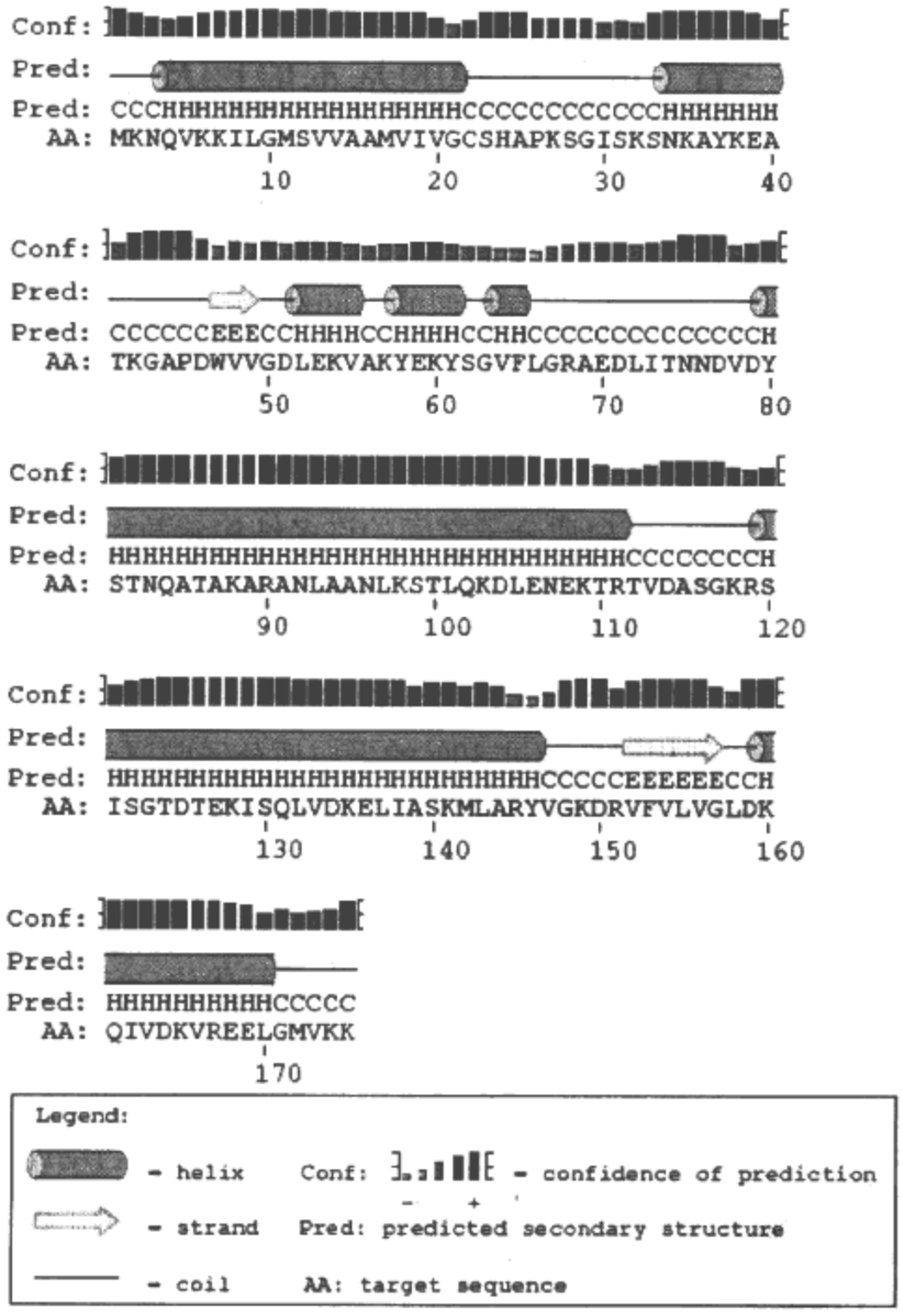


图 2-2　PSIPRED 的蛋白质二级结构预测结果

第三章

利用同源模建法预测蛋白质结构

第一节　同源模建的基本原理及基础

同一个家族的蛋白质成员往往具有类似的结构和功能，如丝氨酸蛋白酶家族、血红蛋白家族等。通过对已经测定的蛋白质的空间结构进行对比，人们发现，蛋白质的三级结构比一级结构更保守，同源蛋白质之间具有相似的空间结构，同源蛋白质之间具有保守的结构内核，差异仅存在分子表面的回折区。一般情况下，如果两个蛋白质的序列同源性不低于50%，那么这两个蛋白中90%的 C^{α} 原子均方根偏差不会超过3 Å[1]；如果以结构叠合后两个蛋白质分子等价 C^{α} 原子之间距离的均方根偏差（RMS）为衡量标准，其RMS不会大于1Å。而同源性低于50%的蛋白质分子之间的许多结构参数如 C^{α} 原子的RMS、残基的溶液可及性、侧链构象等的保守性都会减弱，但是依然具有较强的相似性。甚至对于两个序列相似性为20%左右的远亲蛋白，在某些情况下，其等价的主链原子之间的RMS有时候仍然可以在1.8～3.0Å，相同折叠方式的区域有时候会超过整个蛋白分子的一半以上。

通常情况下，可以将序列相似性超过30%的蛋白质分子归属为一个家族，同一个蛋白质家族的成员由同一个“祖先”进化而来，它们之间具有保守的三维结构。当一个蛋白质家族中的某一个成员测定了三维结构之后，就可以用这个结构作为模板，来预测该家族的其他蛋白成员的三维结构。实际证明，一个蛋白质如果先利用已知的同源蛋白质的结构作模板进行初步预测，然后再利用理论计算方法对模型进行优化，就可以得到可靠性较高的预测结构。

一般来说，模板蛋白和目标蛋白的关系程度越密切，那么得到的结构模型质量就越好。如果模板蛋白和目标蛋白序列的相同率超过了60%，那么就算用自动比对的方法都可以得到很理想的结构模型。如果目标蛋白与模板蛋白序列的残基相同率下降到一定程度，有时候根据实际情况，就需要采取一些干预措施，比如利用多序列比较或更多的手动调整，使序列联配达到最佳，这些干预就成为模建成功必要的补充。

如果一个待预测的序列可以拿到几个模板结构，通常比较合适的做法是先将这些模板结构叠合，然后利用模板中原子的平均位置进行模建。通常情况下，模建时使用多个模板结构会比只用一个模板结构的效果要好，结构保守区的结构也比非保守区域更可靠。

1969年Browne及其同事首先以鸡蛋清溶菌酶的结构为基础，手工模建了α-乳白蛋白的空间结构，并获得了成功，从此开创了利用同源模建技术预测蛋白质空间结构的先河。1981年，Greer等人建立了利用多个同源蛋白质进行结构预测的方法，并将该方法用于模建哺乳动物丝氨酸蛋白酶的结构。

---

[1] 1Å=10^{-10}m。

同源模建是蛋白质三维结构预测的主要手段，也是目前为止蛋白质结构预测中最为成功、准确度最高的方法，但同源模建有一个前提条件，即要模建的目标蛋白必须有一个或多个已知结构的与之序列高度相似的同源蛋白。通常情况下，目标蛋白与模板蛋白的序列一致性不低于30%，如果二者能够进行联配的区域长度超过100个氨基酸残基，序列一致性在25%以上也可以进行同源模建。如果模板蛋白与目标序列的序列一致性在15%～25%，即所谓的“交界区域”（twilight zone）则不能利用简单的同源模建来预测结构。研究表明，序列一致性在交界区域的蛋白质之间是否有同源关系无法根据序列来确定。对于同源性在这个区域的蛋白质进行比较后发现，分子中等价的二级结构单元之间的相对位置偏差可达7Å，方向偏差可达30°。因此在同源模建中如果参考蛋白和目标蛋白之间的序列进化关系较远，那么就会出现很多严重问题。尽管粗略的整体蛋白结构特征可以很好地模建出来，但某些具体的区域可能由于几个残基的联配错误而导致模建错误。模建出来的结构很可能会出现α螺旋之间相对方位的错误或β折叠股之间的错误连接，另外还会造成侧链的错误堆集。在接下去的分子力学或分子动力学优化中，这些错误不但不会被清除掉反而会引入新的结构扭曲。这是因为基于错误的初始结构的能量优化会导致以若干氢键断裂为代价的整个二级结构的重新定位。只有进化上紧密关联的蛋白质分子之间的 C^{α} 原子 RMS 才会小于2Å，所以用于同源模建模板的参考蛋白与目标蛋白之间的序列同源性最低限度不能小于25%，否则模建的结构其可靠性就大打折扣。当然如果有其他生物学证据证实两个序列相似性低的蛋白之间在进化上具有明显的亲缘关系，由同源性较低的蛋白结构模建目标蛋白有时也可能会获得比较好的结果。通常的情况下，如果利用序列相似性搜索无法寻找到同源性高于25%的模板蛋白，则需要利用折叠模式识别的方法来寻找参考蛋白。

第二节　同源模建预测蛋白质结构的基本步骤

同源模建过程主要可分以下几个阶段。

一、参考蛋白搜索

根据目标蛋白的一级序列搜索参考蛋白是同源模建的第一步，搜索策略的掌握直接关系到以后模建的成功与否。搜索的范围为蛋白质结构库（PDB）。

二、确定结构保守区

如果目标蛋白有两个以上已知结构的参考蛋白，可通过这些参考蛋白之间的结构叠合来确定结构保守区（SCRs），这是比较理想的情况。如果参考蛋白中只有一个有空间结构，那么结构保守区的确定就必须通过多重序列对比的方法来实现。

三、蛋白主链的模建

结构保守区的主链坐标可以直接由参考蛋白拷贝下来，主链的模建关键在于环区。一般环区的模建有两个途径。一种途径是片段搜索，根据待模建环区的端点距离及环区两端结构保守区的主链构象在已知的片段库里搜索出若干（一般是 10 个）最符合条件的片段供参考。另一种环区模建的途径是自动生成法。自动生成法不需要任何模板片段，它利用随机搜索的方法确定环区构象。这种方法适合很小环区的模建，对于稍微大一些的环区耗时太长而且随机搜索的构象空间也不一定完全，因而结果不一定好。

四、侧链安装

统计研究表明，蛋白质分子中残基的侧链取向不是杂乱无章的，事实上一个氨基酸残基的侧链构象可以归结为几种分立的最可几构象，这些分立的构象形成一个转子文库。侧链搜索的任务就是在转子文库中挑选出最佳的侧链构象组合。

五、优化处理

蛋白质分子的主、侧链都确定后只能说得到了一个初步结构，需要进一步优化使之更加合理。优化的手段主要是分子力学和分子动力学。

六、合理性检测

模建的结构合理与否可以通过多种不同的方式来检验，最常用的是 Profiles-3D 检测。Profiles-3D 可以检测出模建的蛋白中哪些残基的构象不合理，不合理程度多大，这样研究者可以依此对不合理的部分再优化或改变策略重新模建。

第三节　利用 SWISS-MODEL 服务器进行蛋白质结构同源模建操作实例

蛋白质同源模建技术逐渐成为一个较为常规的分析方法，目前互联网上提供了很多利用同源模建技术预测的蛋白质结构的服务器，比如 SDSC1（San Diego）、3D-JIGSAW（London）、FAMS（Tokyo）和 SWISS-MODEL（Geneva）。其中 Guex 等人发展的 SWISS-MODEL 是使用最广泛的网络模建服务器，SWISS-MODEL 是一个自动化的、对学术团体免费的蛋白质结构模拟环境，任何人都可以登录网站进行蛋白质结构的同源模建。本节以鼠尿酸氧化酶的结构模建为例说明如何利用 SWISS-MODEL 服务器来进行蛋白质的三维结构模建。

一、登录服务器，输入目标序列

可以通过 http://swissmodel.expasy.org 访问 SWISS-MODEL 服务器的 web 界面，如图 3-1。

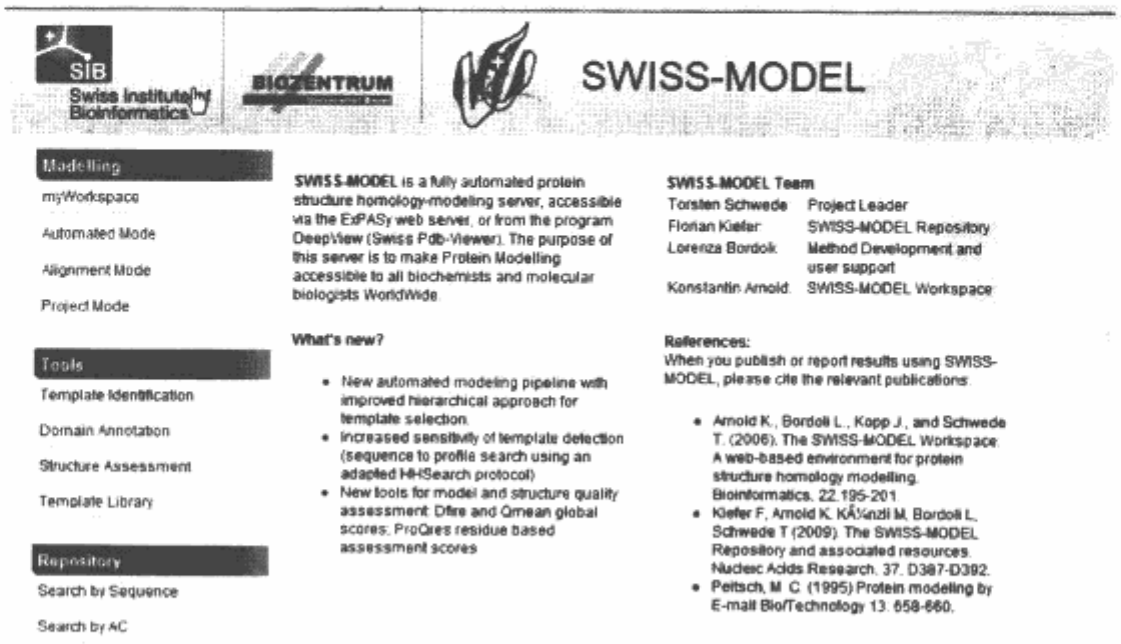


图 3-1　SWISS-MODEL 服务器初始页面

自动模式（Automated Mode）：用于建模的氨基酸序列或是 Swiss-Prot/TrEMBL 编目号（accession）可以直接通过 web 界面提交。用户可以选择指定模板结构，模板可以来自由 PDB 数据库抽取得到的内建模板库，也可以上传 PDB 格式的坐标文件。

联配模式（Alignment Mode）：这个模式需要多序列联配的结果，联配的序列中需要包括目标序列和模板。用户需要指明联配文件中哪一条序列是目标序列，哪一条作为模板。服务器会基于比对结果建模。

项目模式（Project Mode）：这种模式允许用户提交经过手工优化的请求给服务器。DeepView 被用来建立一个项目文件，它包含了模板结构，以及目标序列与模板的联配结果。这个结果也要上传到服务器。这种方式提供对建模过程中细节的控制，例如选择不同的模板，手工编辑目标序列和模板的联配结果，以便正确地定下插入和删除的位置。项目模式还能够用于重复改进 First Approach mode 的结果。

选择自动模式。在本实例中，选择 Automated Mode（自动模式），点击 Automated Mode 键之后，就会进入自动模建模式页面（图 3-2）。需要在 Provide a protein sequence or a UniProt AC Code 对话框填入需要预测结构的蛋白质序列，序列可以填入 FASTA 格式或填入序列在 UnitProt 数据库中的登录号也可以。

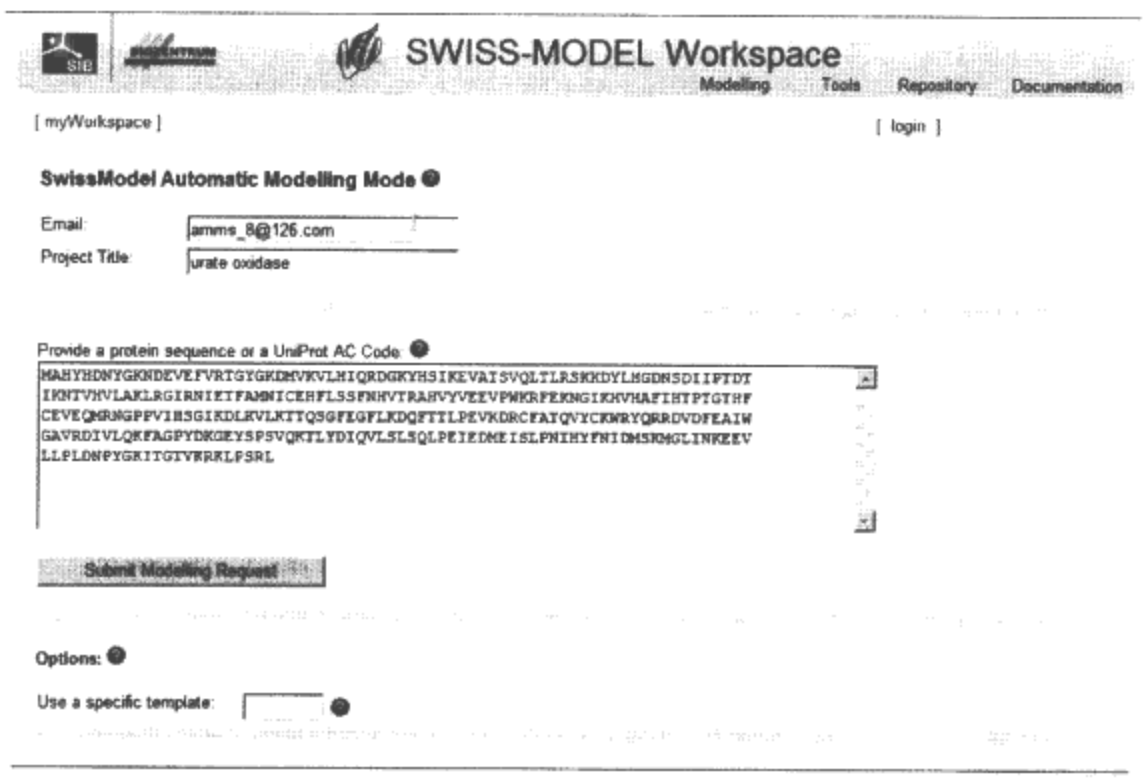


图 3-2　SWISS-MODEL 服务器命令界面

在本实例中，直接把鼠尿酸氧化酶的氨基酸序列粘贴在序列提交对话框；使用者的电子信箱可以填写也可以不填写，如果填写使用者的电子信箱，提交任务后，使用者会收到邮件提示。填写完模建信息之后，点击 Submit Modelling Request 键，即可向模建服务器提交模建任务。提交任务后，模建页面会显示任务提交成功。同时，因为填写了 email 地址，因此在填入的在信箱中收到了提交任务成功的邮件。

二、模建结果与分析

提交任务大约几分钟之后，邮箱会收到模建任务已经完成的邮件。同时，SWISS-MODEL 服务器页面也会自动更新，显示模建结果（图 3-3）。在本实例中，模建的是鼠尿酸氧化酶的结构，模建结果中模建的范围是 18～297，模建采用的模板其分辨率是 2.8Å，

目标序列与模板序列的相似性是 34.576%。模建获得的结构模型可以通过点击 DeepView 查看，另外也可以点击 download 键下载为 pdb 格式的文件。可以通过模建结果网页中的 Print/Save this page as 将结果文件保存在一个 PDF 文件中。

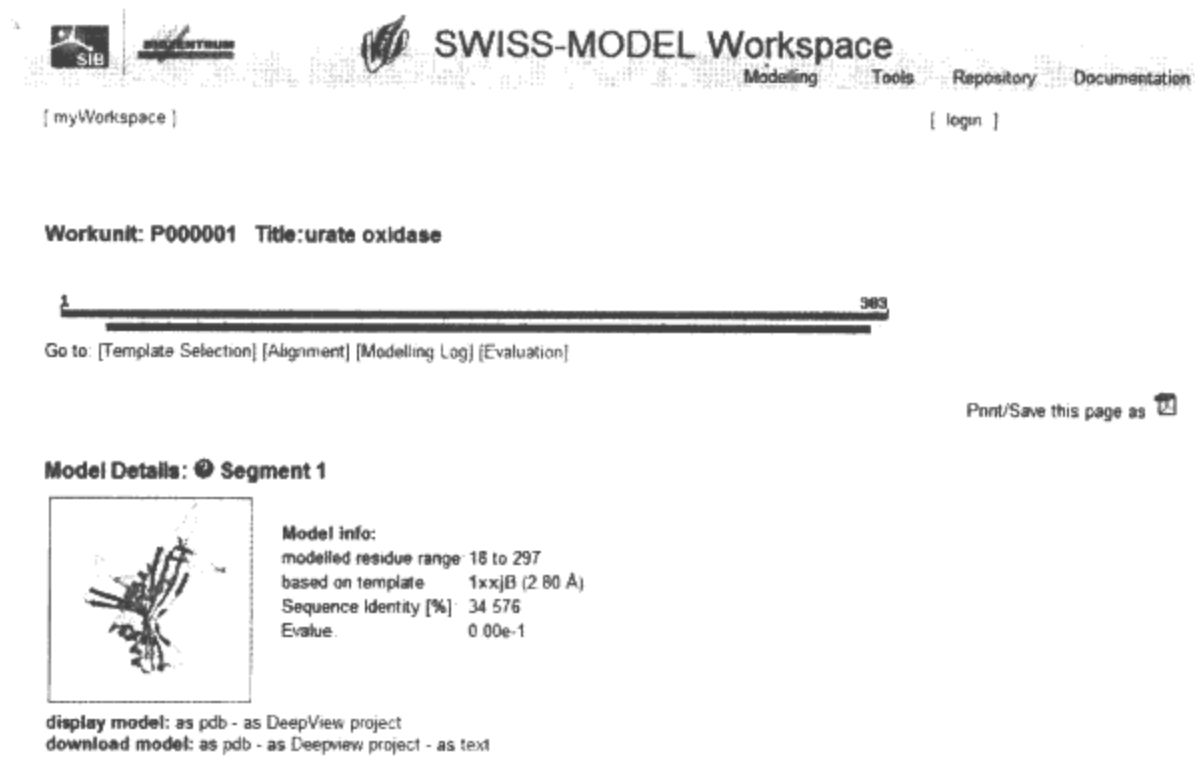


图 3-3 SWISS-MODEL 服务器的模建结果信息

在弹出的模建结果网页中，还有目标蛋白序列与模板结构的序列联配信息（图 3-4）、对模建结构的合理性评估信息（图 3-5）。

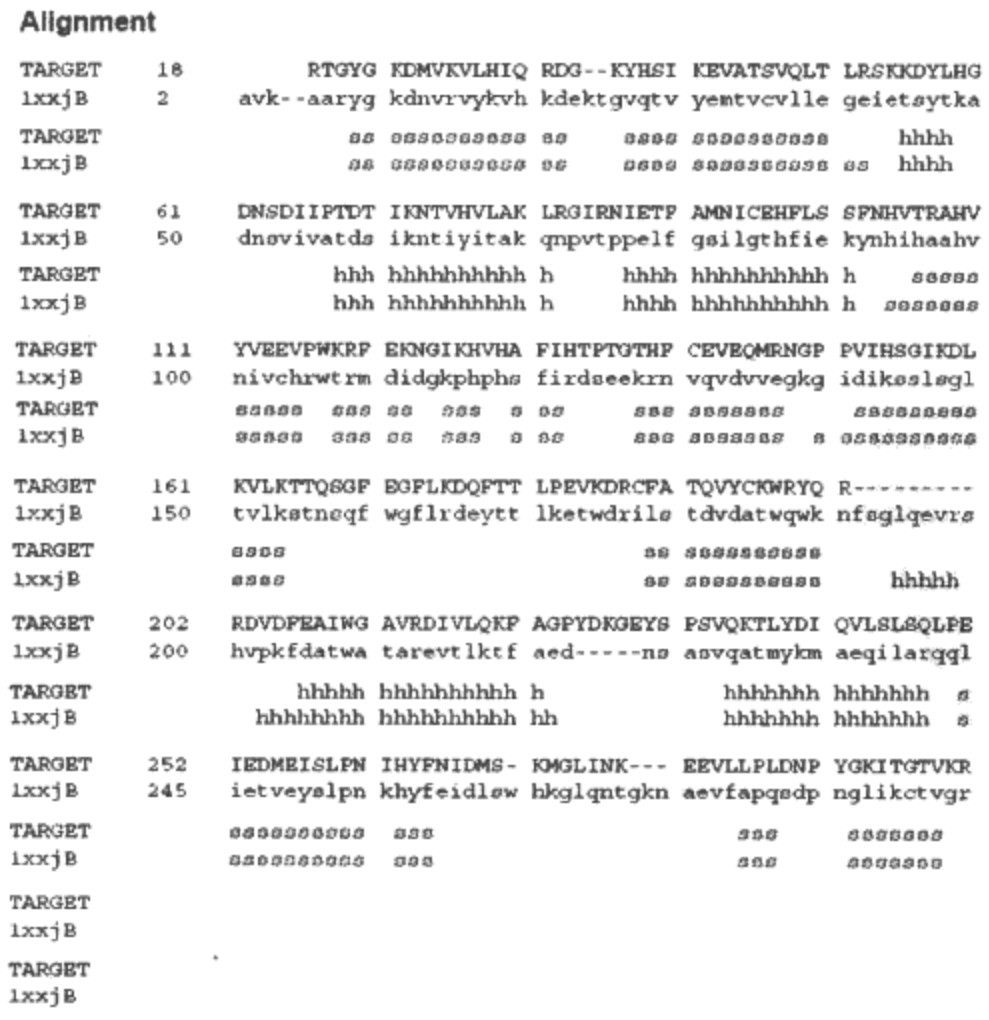


图 3-4 SWISS-MODEL 模建过程，目标序列与模板的序列联配

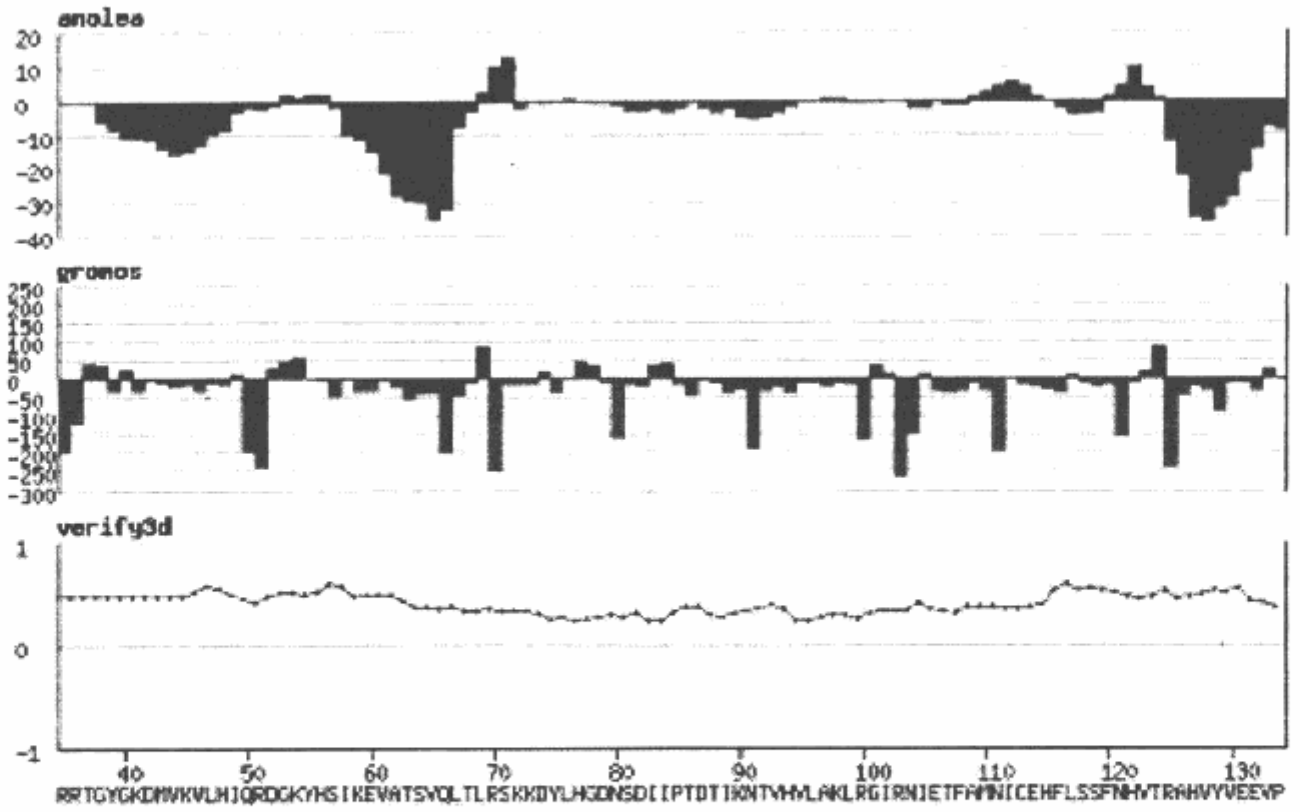


图 3-5　SWISS-MODEL 服务器提供的对模建结果的评估（部分）

联配信息中包括目标序列与模板序列的比对信息及二级结构分析情况。SWISS-MODEL 服务器结构合理性评估有三种方式，分别是原子间经验平均力势函数 Anolea、Gromos 分子动力学以及表示蛋白质残基与结构环境匹配程度的 Verify3D 评估。可以通过点击评估方式后面的 on 及 off 键来决定选择那种评估方式。本节中选择了全部三种评估方式。

此外，还有模建产生的日志文件，主要介绍模板搜索的过程说明等信息。限于篇幅，这里不一一列出，读者可以直接通过 SWISS-MODEL 服务器预测之后仔细查看。

如果提交的序列没有搜索到合适的模板（与目标序列的序列相似性大于 25%），服务器也会显示搜索不到模板的提示。

利用 SWISS-MODEL 执行蛋白质结构的同源模拟，操作方便容易，但是它仍然有一定的局限性。

① 优化不足。SWISS-MODEL 执行结构模建时，为了避免因为得到最低的总能量而使结构的改变太大，因而在进行能量最小化过程时，可能会使能量最小化的步调（steps）减少，相对地结构可能尚未达到能量稳定的状态。

② 假如 target 与模版之间序列的一致性必须达到 25%以上，如果达不到此要求时，结构搭建无法进行。

③ 搭建的 Target 最终结构的长度会受限于模版结构的长度，Target 序列比模板序列中多的氨基酸片段的结构不会自动模建。

第四节　利用 Discovery Studio 结构模拟系统进行同源模建操作实例

Accelrys 公司的 Discovery Studio™ (简称 DS)是基于 Windows/Linux 系统、面向生命科学领域的新一代分子建模和模拟环境。Discovery Studio 从原来的 Insight Ⅱ 三维图形环境软件包开发而来。Discovery Studio 针对生命科学应用，提供生物分子及有机小分子建模的

显示工具、功能分析工具、结构改造工具、动力学模拟工具等，帮助研究人员在实验前全面了解生物分子的结构与功能，从而有针对性地设计实验方案，提高实验效率，降低科研成本。Discovery Studio 在揭示蛋白质结构功能关系、生物分子结构模拟与动力学计算、基于靶点药物设计、抗体设计、酶工程、生物分子间的相互作用（包括蛋白质与蛋白质、蛋白质与肽、蛋白质与核酸、蛋白质与有机小分子）、生物分子核磁共振、功能基因组以及蛋白质组等方面有着广泛的应用。DS 目前的主要功能包括：蛋白质的表征（包括蛋白-蛋白相互作用）、同源建模、分子力学计算和分子动力学模拟、基于结构药物设计工具（包括配体-蛋白质相互作用、全新药物设计和分子对接）、基于小分子的药物设计工具（包括定量构效关系、药效团、数据库筛选、ADME/T）和组合库的设计与分析等。

Discovery Studio 为用户提供了一整套利用 Homology Modeling 方法自动预测蛋白质空间结构的工具。用户只需要提供蛋白质的氨基酸序列就可以轻松完成模型构建及模型可信度评估的工作。DS 的 Homology Modeling 主要基于 MODELER 程序。目前 MODELER 已成为学术上使用最为广泛、预测最为准确的同源建模工具之一。其主要的建模步骤包括：

① 使用序列相似性工具 BLAST 或 PSI-BLAST 搜寻目标序列的模板；

② 使用结构比对方法将模板进行比对，叠合；

③ 使用序列比对方法将目标序列与模板结构的序列进行比对；

④ 使用 MODELLER 产生目标序列的模型；

⑤ 模型的评估。

本实例中使用一个胞外淀粉酶的模型构建过程为例子，展示如何使用 DS 为该淀粉酶自动构建空间结构，并对所构建的模型进行评估。

一、识别模板，比对模板

本实例使用 BLAST 来搜索 templates。进行 BLAST 搜索时，数据库通常使用 Protein Data Bank（PDB）数据库。

1. 载入序列

从 Files 中，打开 Samples | Tutorials | Protein Modeling | P41131.fasta（图 3-6）。

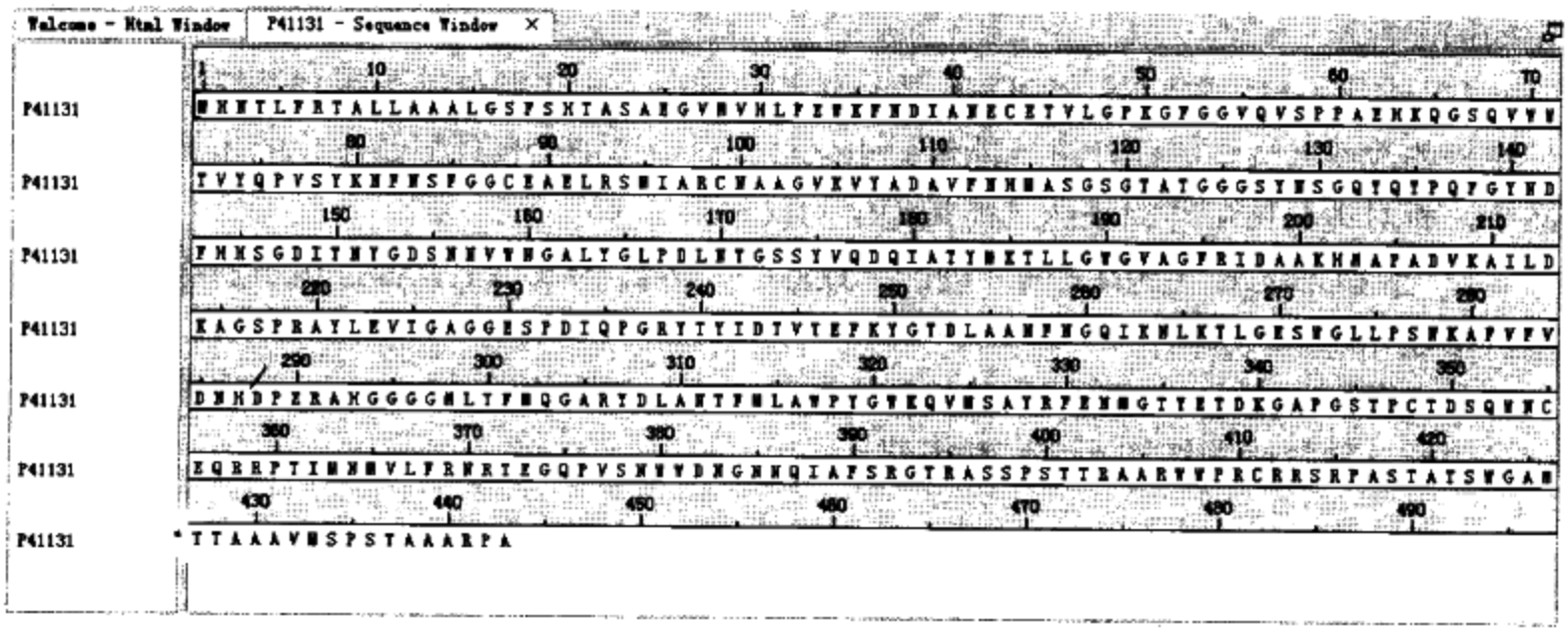


图 3-6　从 Files 读入序列

2. BALST PDB_nr95 数据库，寻找模板

（1）选择 target　在 Protocols 中，展开 Sequence Analysis 文件夹，双击 BLAST Search

(DS Server)。在 the Parameters 中，点击 the Input Sequence parameter，选择 P41131:P41131 Input Sequence 中的文件名为 sequence window 的名字（P41131）与该窗口中的序列名称（P41131）的名字组合。

（2）选择 BLAST 数据库

Input Database 选择 PDB_nr95（图 3-7）

注意：PDB_nr95 序列数据库已经安装在 DS server 上。如果需要 BALST 其他数据库，用户需要另外安装相应的数据库。

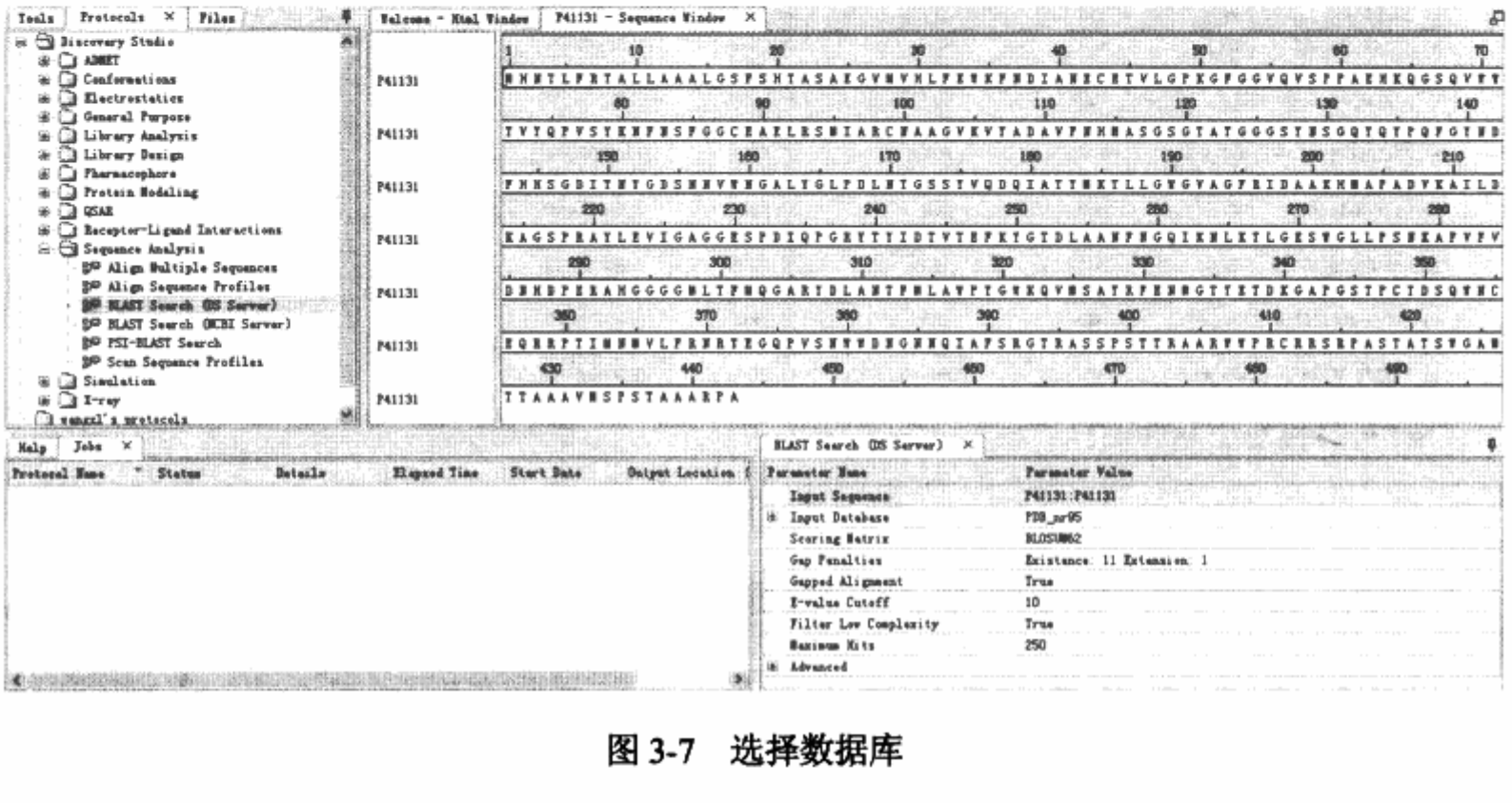


图 3-7　选择数据库

（3）运行 Protocol

在 Protocols toolbar，点击 ▶ 运行，等待计算完成

计算完成后，会显示一个 Job Completed 对话框（图 3-8）。点击 OK。

（4）查看计算结果

在 Jobs（作业）中，双击 the completed job

这将打开一个 Html 的窗口（图 3-9），里面包含 Reprot.htm 文件（该文件为比对结果报告）。

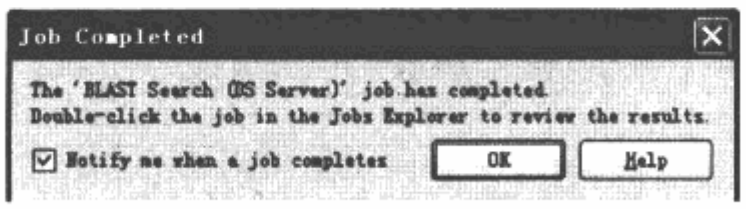


图 3-8　Job Completed 的对话框

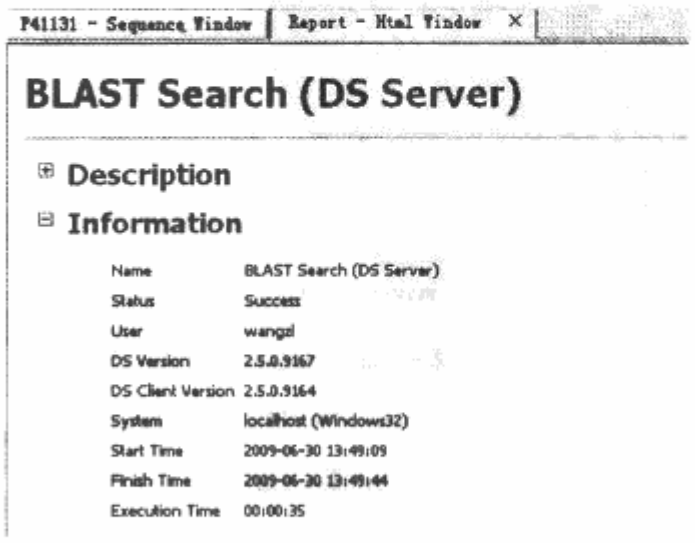


图 3-9　结果报告文件

Html 窗口中，Output Files 部分，点击 the View Results。这将打开 BLAST 搜索找到的序列。

注意：由于没有设置结果的保存路径，BLAST 的结果保存于默认文件夹 My Documents\Discovery Studio Client\Results\BLASTSearchDSServer_<time stamp><ID>。

在 P41131 - Blast 窗口，点击该窗口下的 Table View。Table View 显示了命中的序列（图 3-10）。每行表示一条命中的氨基酸序列。在 DS 中，灰色的 cell 不能被编辑。

P41131 - Sequence Window | Report - Html Window | P41131 - Blast Window ×

	Title/Descripti…	Accession	Sequence Length	Alignment Lengt…	Bit Score	E-value	Identity	Positive	Resolutio…
1	ALPHA-AMYLASE…	1G94_A	448	381	317.39	2.30147e-87	47	58	1.74
2	Alpha-amylase…	3DHP_A	496	424	300.827	2.21063e-82	44	53	1.5
3	ALPHA AMYLASE…	1HXO_A	496	425	291.582	1.51995e-79	44	53	1.38
4	ALPHA-AMYLASE…	1JAE_A	471	405	286.96	3.44408e-78	41	52	1.65
5	alpha-amylase…	3DCO_A	422	188	88.1965	2.55228e-18	30	45	2.78
6	Alpha-amylase…	1UA7_A	422	188	86.6557	7.18225e-18	30	45	2.21
7	CYCLODEXTRIN …	1D3C_A	686	284	55.6398	1.20798e-08	23	37	1.78
8	CYCLODEXTRIN …	1PAM_A	686	284	54.8842	2.75931e-08	24	35	1.8
9	Cyclomaltodex…	3BMV_A	683	266	54.299	3.54413e-08	23	38	1.6
10	CYCLODEXTRIN …	1CGT_A	684	231	46.595	7.1474e-06	22	37	2
11	4-ALPHA-GLUCA…	1LBJ_A	441	146	46.2098	1.02321e-05	27	39	2.5
12	CYCLODEXTRIN …	1CTG_A	680	147	44.2838	3.88819e-05	24	41	2.5
13	ALPHA-AMYLASE…	1HVX_A	483	106	44.2838	4.48074e-05	31	42	2
14	ALPHA-AMYLASE…	1HT6_A	404	230	42.3578	0.000148524	24	36	1.5
15	BARLEY ALPHA-…	1AVA_A	403	255	41.9726	0.000183548	25	41	1.9

Map View | Table View | Text View

图 3-10　Table View

注意：命中的序列按照 E 值进行升序排序。E 值最低的序列，结果最可靠，排在第一行。

点击 Map View tab

Map View 将命中结果都显示在一张图中，每条线表示一条序列（图 3-11）。每根横条根据打分不同而配以不同的颜色（分数超过 400 为红色，最佳的命中结果）。本例中的 target P41131 放在窗口的最上方，为一条长度为 443 个氨基酸的直线。

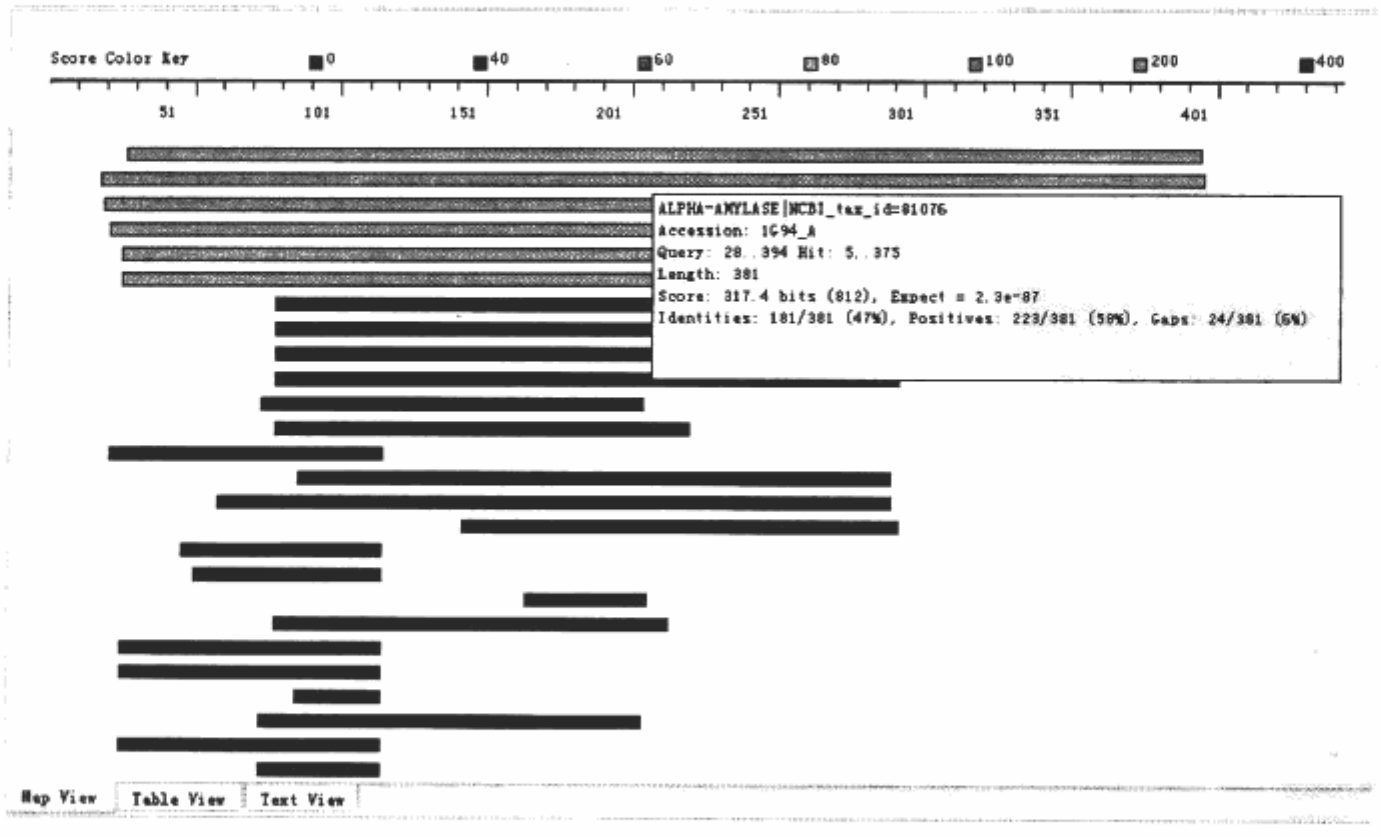


图 3-11　Map View

用户可以将鼠标放置在某一个命中序列上，如下信息将会显示（如图 3-11）：

- 序列数据库的描述
- 序列的编号
- 目标序列中的起始氨基酸位置
- 数据库中命中序列的起始氨基酸位置
- 命中序列的长度
- 命中序列的分数

滑动鼠标的中间键可以放大（图 3-12）（缩小）Map View 中的结果

这时用户可以看到窗口顶端 target 的相应氨基酸。可能需要放大几次才能看见具体的氨基酸类型。

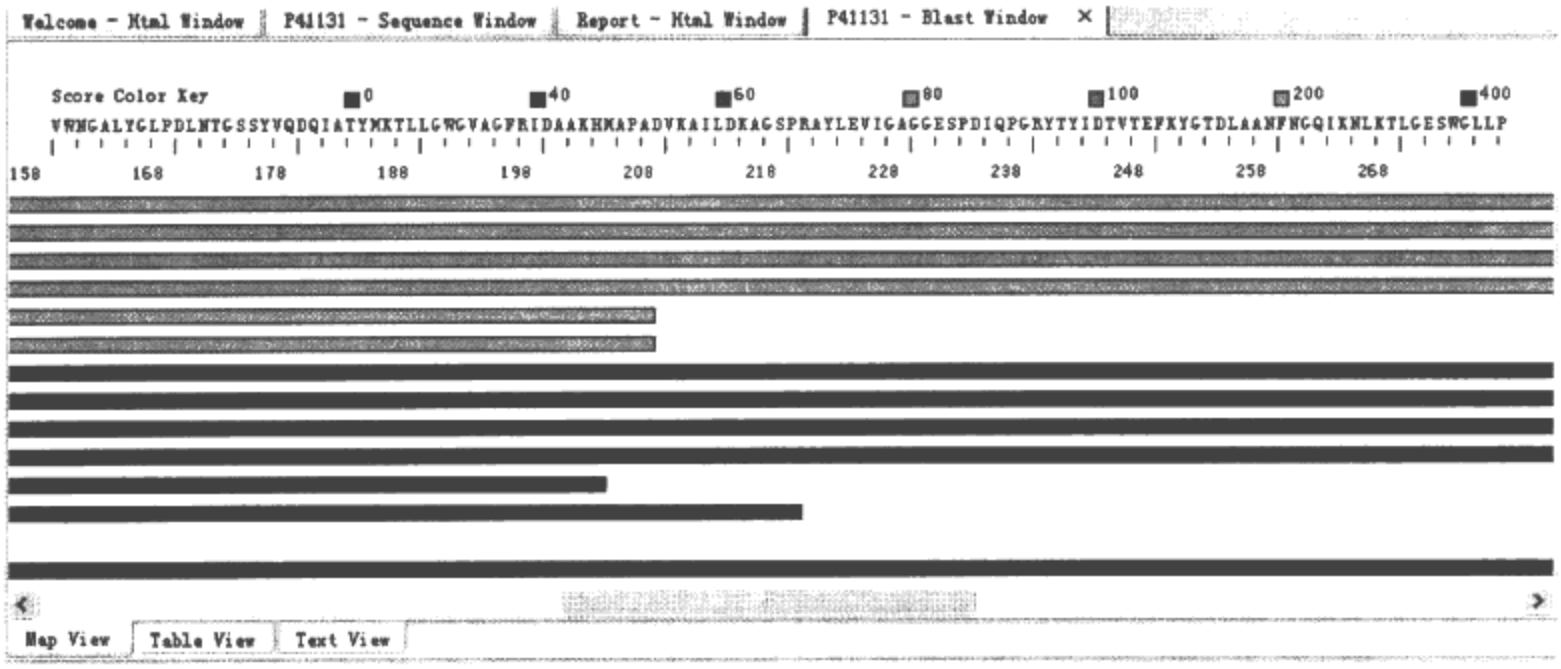


图 3-12　放大的 Map View

注意：Map view 中命中序列的顺序并没有改变。

二、将模板进行比对

为了构建 target 的 3D 结构，需要挑选一个或多个合适的同源模板（templates）。一个理想的 template 需要涵盖整个 target 的长度，具有较高的序列等同性（Sequence identity），并且 E 值要够小（$<1\times10^{-5}$）。根据上述原则，选用前 4 个命中序列作为 template。

一般而言，若有多条模板（模板之间相似度不能太高）与 target 具有相似的同源性，那么都将使用多模板来构建同源模型。这些模板在核心区域一般都高度保守，而在一些 loop 区的构象上则有所不同。target 可以与其中某个模板在某 loop 区匹配，而与另一个模板在另一 loop 区匹配。因此，使用多模板可使建模过程中模型的每个部分都能找到最合适的模板。

1．载入模板结构及其与 target 的比对结果

点击 P41131-Blast 窗口，点击 Map View，按住 SHIFT 键同时点击前四个命中序列 1G94_A、3DHP_A、1HX0_A、1JAE_A，将其选中，右键点击，选择 Load Structure and Alignment。这将在 DS 中打开两个新的窗口。一个为新的 3D 窗口，记为 1G94A，包含了上述四个模板结构的 A 链以及结构中的水分子和配体分子（图 3-13）；另一个为序列比对窗口（记为 Sequence），包含了目标序列 P41141 和四个模板结构的序列比对图（图 3-14）。

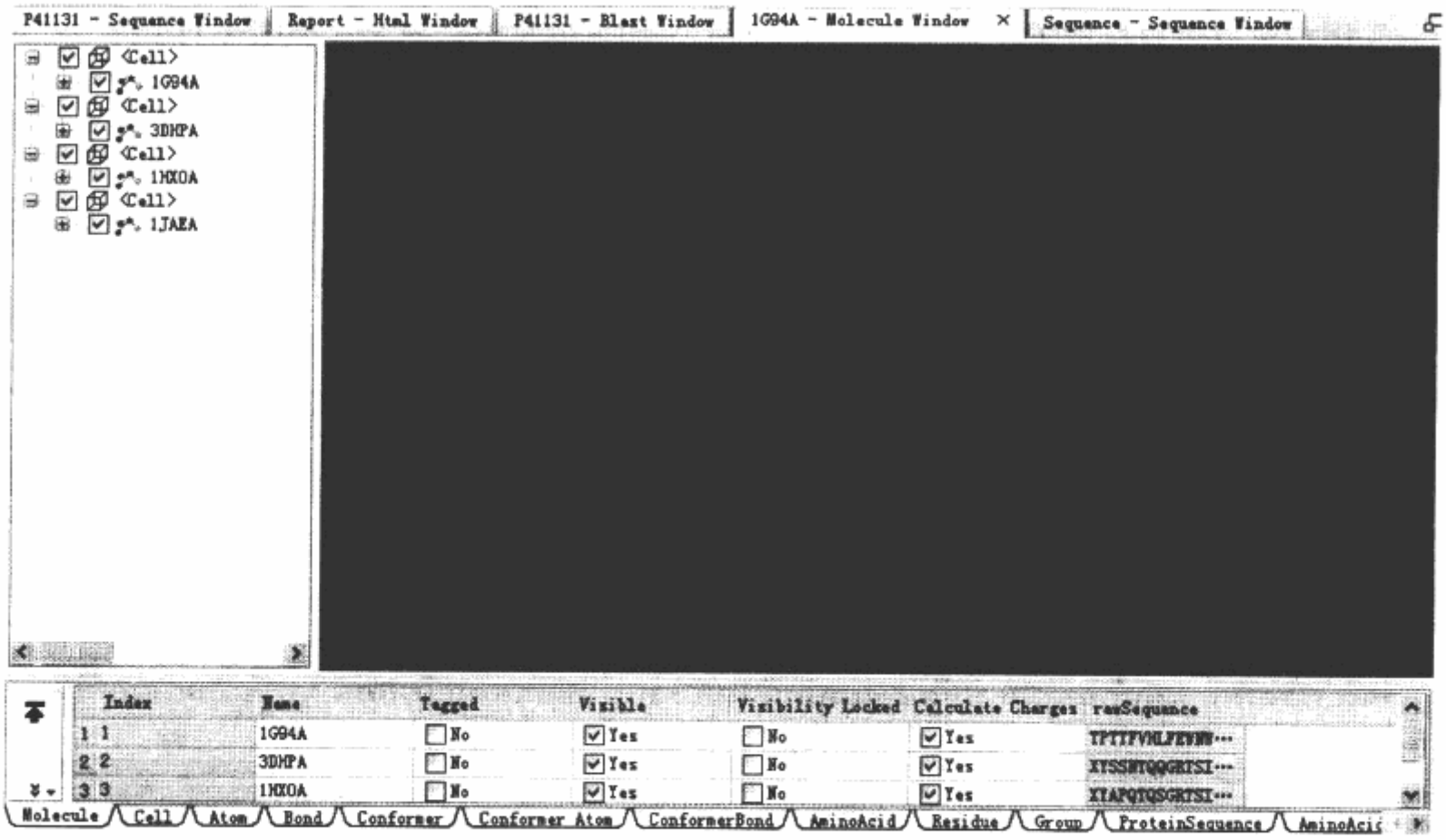


图 3-13　模板结构窗口

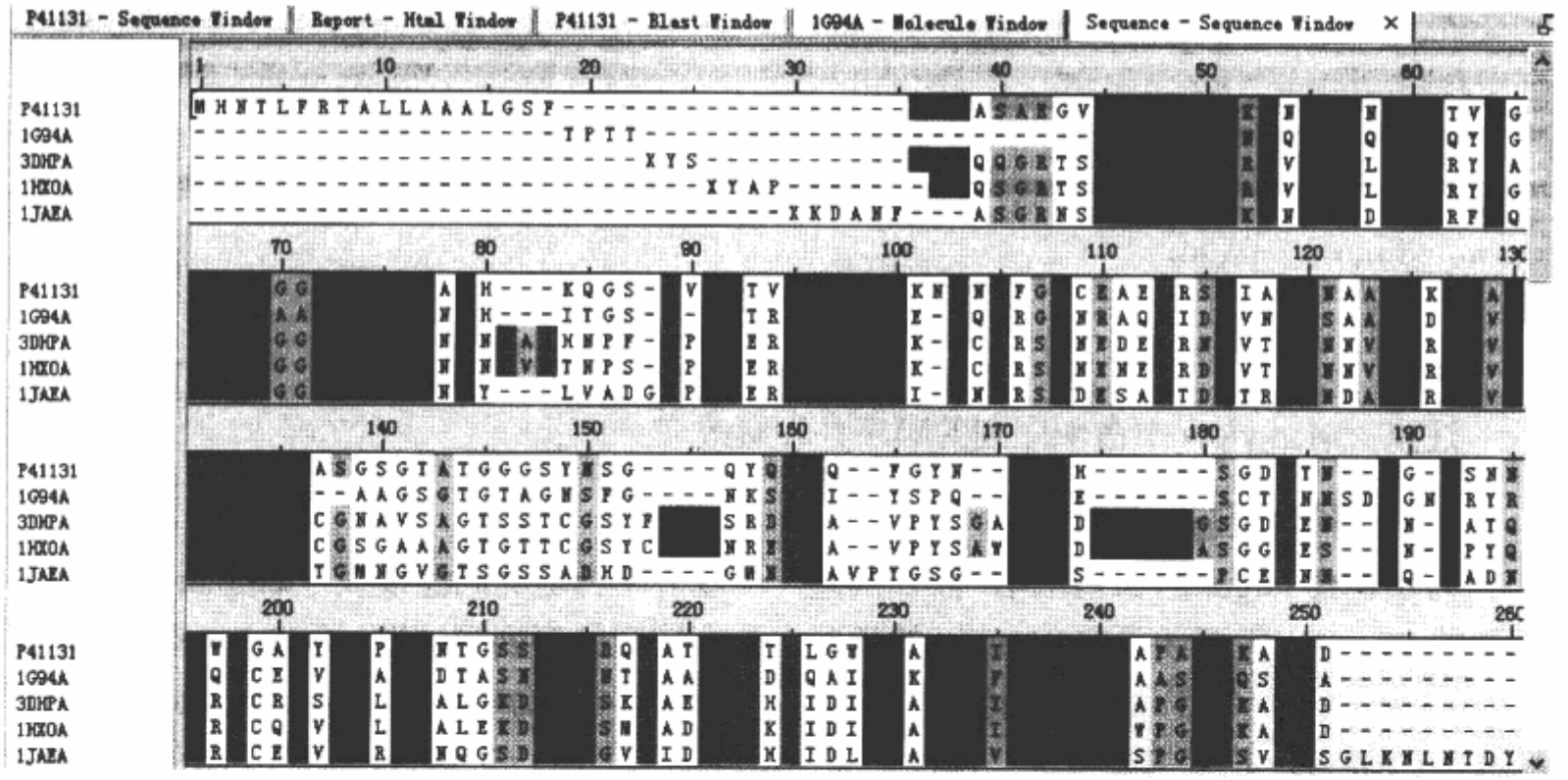


图 3-14　序列窗口

2．查看二级结构和 BLAST 比对结果

在 Sequence 中点击右键并选择 Secondary Structure Cartoon

四个模板序列的下方会出现根据 PDB 和 Kabsch-Sander 两种分类方法进行分类的该序列的二级结构图。红色横条表示 α 螺旋，蓝色箭头表示 β 折叠（图 3-15）。

检查目标序列与模板间的比对结果。发现 C 端中约有 50 个氨基酸没有比对上。这样的比对结果不适合用于同源建模，需对序列进行重新比对。

3．模板结构间的结构比对

序列保守性和结构保守性通常来说有所不同，所以序列比对的结果常常与结构比对的

结果不同。对于同源建模来说，在模板与 target 进行比对之前，最好能将模板基于其结构的相似性先进行一次比对。

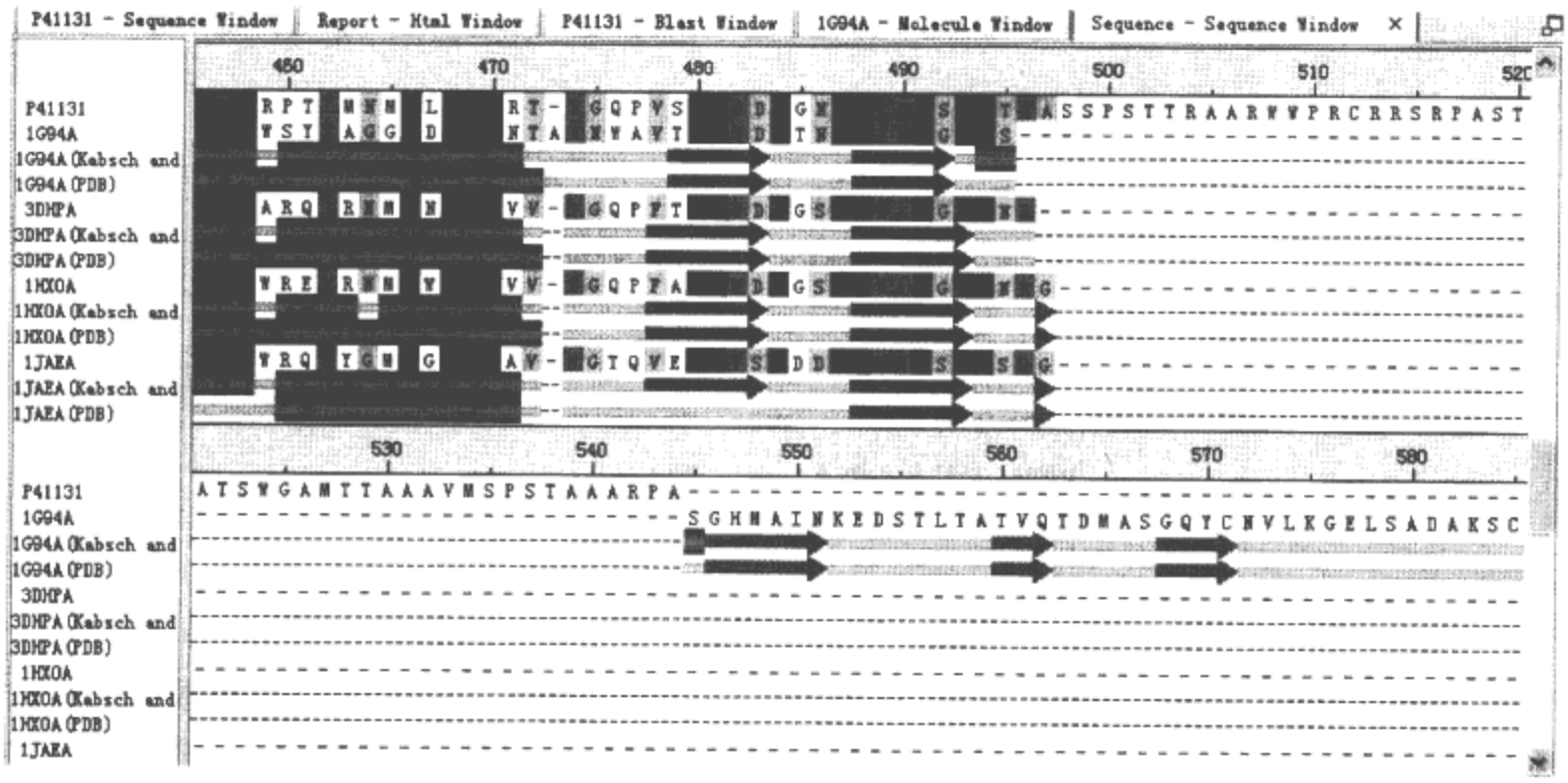


图 3-15　模板二级结构分布图

在 Protocols 中，展开 Protein Modeling 文件夹，双击 Align Structures (MODELER)。在参数 Parameters 中，单击 Input Sequence Alignment 并选择 Sequence。展开 Input Sequence Alignment（点击前面的“+”号），可以观察到四个蛋白质结构 1G94A、3DHPA、1HX0A、1JAEA 自动填充到 Input Protein Structures 里面。

BLAST Search (DS Server)　Align Structures (MODELER) ×

Parameter Name	Parameter Value
Alignment Type	Align Structures
⊟ Input Sequence Alignment	Sequence
Input Protein Molecules	1G94A, 3DHPA, 1HXOA, 1JAEA
⊞ Reference Sequence Alignment	
Gap Open Penalty	0.0
Gap Extension Penalty	3.0
Overhang	10
Fast Alignment	False
Initialize Alignment	True

在 Protocols toolbar 中，点击 ▶ 运行，等待计算完成。

在 Jobs 中，双击 completed job

结果文件 Report.htm 将显示在一个新的 Html 窗口里。

4．查看结构比对结果

在该 Html 窗口中，滚动页面至 Summary 部分，可以查看每对结构之间两两比对时的主链 RMSD 值以及比对的氨基酸数量。3DHPA 与 1HX0A 非常相似，两者之间比对了 496 个氨基酸，RMSD 值小于 0.5 Å。1G94A 和 1JAEA 与其他蛋白则不那么相似（图 3-16）。

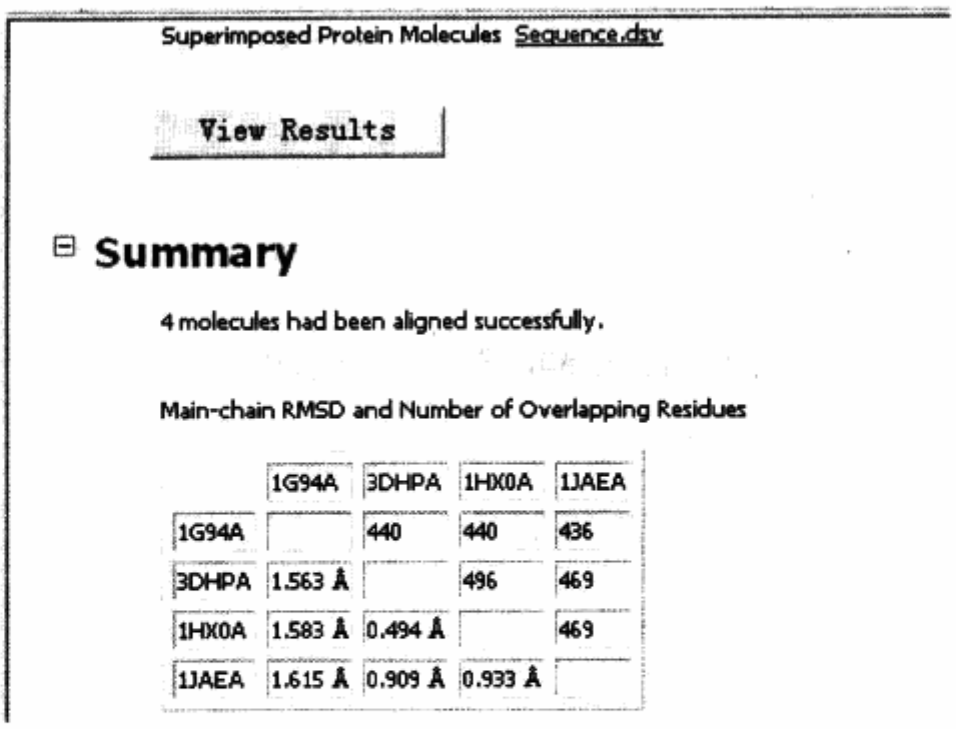


图 3-16　结构比较结果

点击 View Results。这将打开两个新的窗口。一个为序列窗口，里面有模板序列的比对结果（图 3-17）。另一个为 3D 窗口，叠合后的模板分子都在该窗口中。所有叠合的结构都只以 C^{α} 的棒状（stick）形式显示出来（图 3-18）。观察这些结构的叠合找出它们不同的地方。

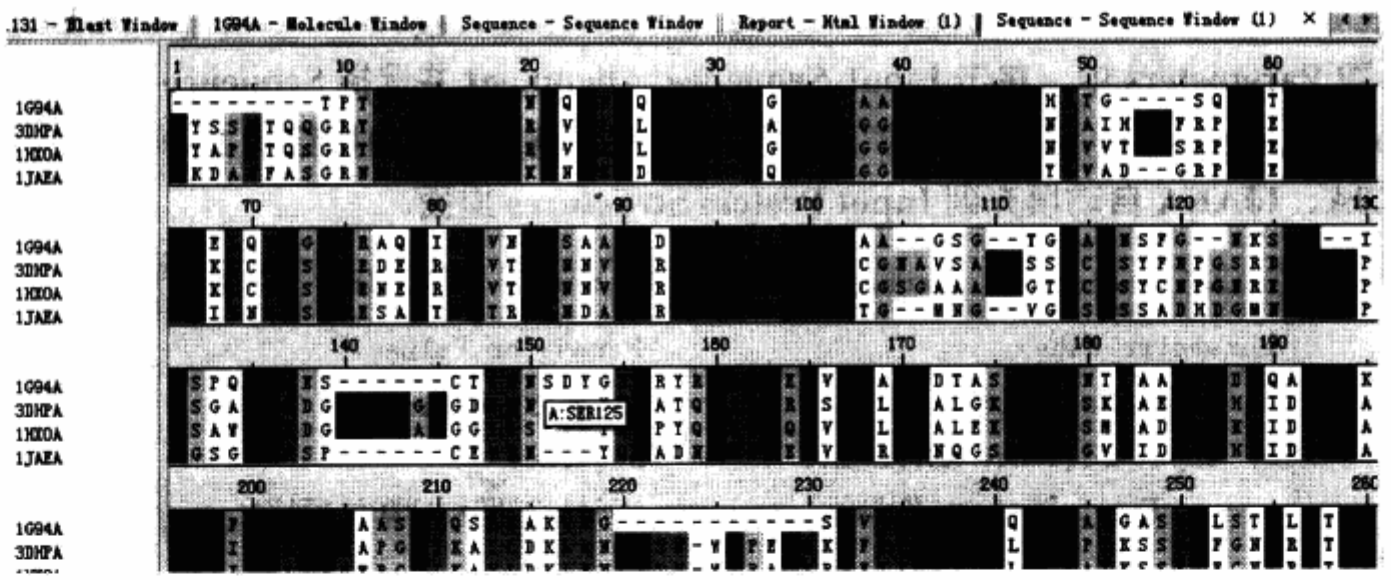


图 3-17　模板序列的比对结果

图 3-18　模板叠合图

三、将目标序列与模板比对

同源建模中，根据目标序列与模板的相似性可以选用不同的比对方法将目标序列与模板序列进行比对：当模板与目标序列的同源性很高时（序列等同性超过 50%），BLAST 可以清晰地识别正确的模板，target 与模板序列间使用简单的序列比对就能获得很好的比对效果；当序列相似性不高但仍高于 twilight 区（序列等同性为 25%～50%）时，虽然 BLAST 还是能够识别出正确的模板，但是，简单的多序列比对已不再能够产生正确的比对结果。可以通过生成序列 profile 的方法来改进序列比对结果；若序列相似性低于 25%，必须使用 PSI-BLAST 来识别模板，而且必须使用序列 profile 来比对 target 和 template。

由于本实例中的 target 与 template 的序列等同性为 25%～60%（最好的 template 为 47%），所以采用第二种比对方法。一个好的序列 profile 必须包含大量非冗余的同源序列比对结果。

本实例中，分三步将 target 与模板进行比对：① 利用 BLAST 搜索 UniRef90 数据库，寻找 target 的同源序列；② 利用多重序列比对将 target 与命中序列进行比对，产生第一个序列 profile；③ 将步骤 2 中产生的 profile 与模板比对所产生的 profile 再进行比对。

1．载入 target

从 Files，打开 Samples | Tutorials | Protein Modeling | P41131.fasta

序列 P41131 在 sequence 窗口中打开。现在，需要用 BLAST Search（DS Server）Protocol 来搜索 UniRef 90 数据库。

2．设置 Protocol 参数并运行 Protocol

在 Protocols 中，展开 Sequence Analysis 文件夹，双击 BLAST Search (DS Server) protocol。提示关闭已经打开的 protocol 时，点击 Yes。

点击 Input Database，选择 UniRef 90。

将 E-value Cutoff 更改为 0.0001，将 Maximum Hits 更改为 500。

BLAST Search (DS Server) ×

Parameter Name	Parameter Value
Input Sequence	P41131:P41131
⊞ Input Database	UniRef90
Scoring Matrix	BLOSUM62
Gap Penalties	Existence: 11 Extension: 1
Gapped Alignment	True
E-value Cutoff	0.0001
Filter Low Complexity	True
Maximum Hits	500
⊞ Advanced	

在 Protocols toolbar 中，点击 ▶ 运行，等待计算完成

计算完成后，在 Jobs 中，双击完成的工作

这将使结果文件 Report.htm 显示在一个新的 Html 窗口里。

3．查看结果

在该 Html 窗口中，Output 文件部分，点击 P41131_profile.pir，这将打开一个新的序列窗口，里面是 BLAST 搜索 UniRef 90 数据库的结果（图 3-19）。

4．将 target 与 BLAST 命中序列重新进行多序列比对

在 Protocols 中，展开 Sequence Analysis 文件夹，双击 Align Multiple Sequences protocol，点击 Input Sequence Set parameter，选择 P41131_profile（图 3-20）。

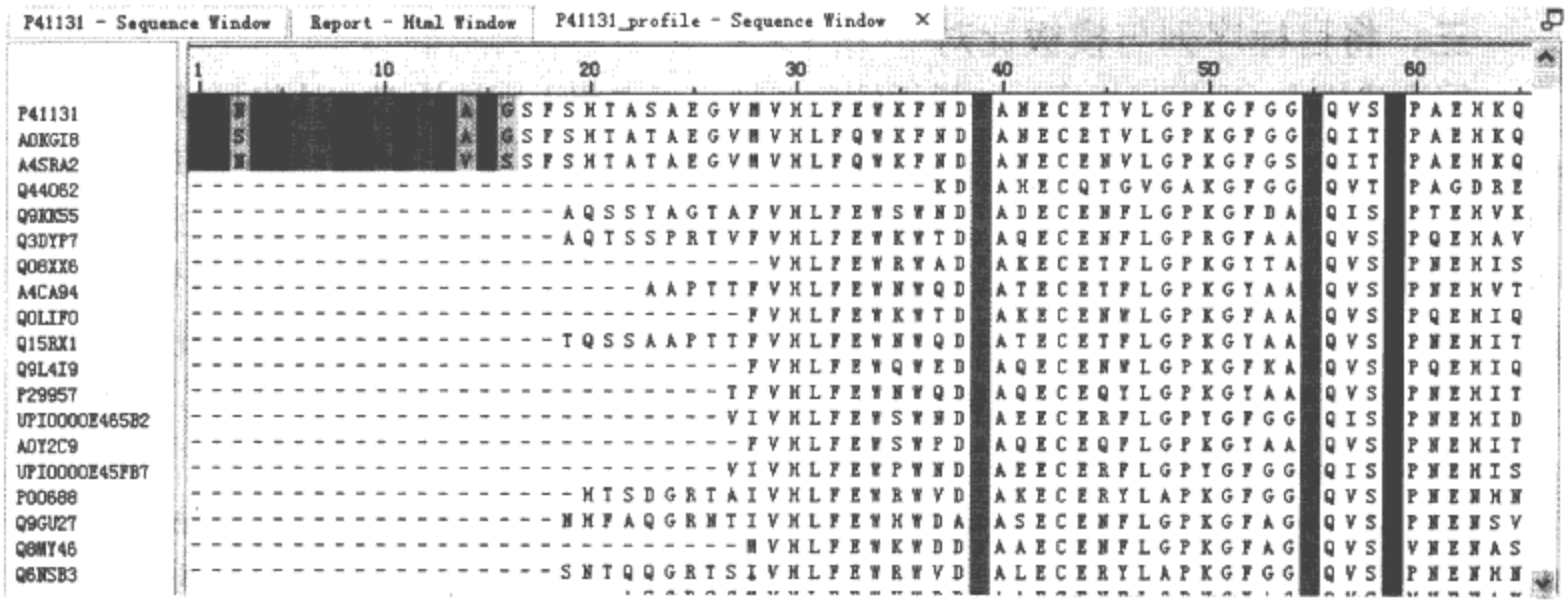


图 3-19　BLAST 搜索 UniRef90 数据库的结果

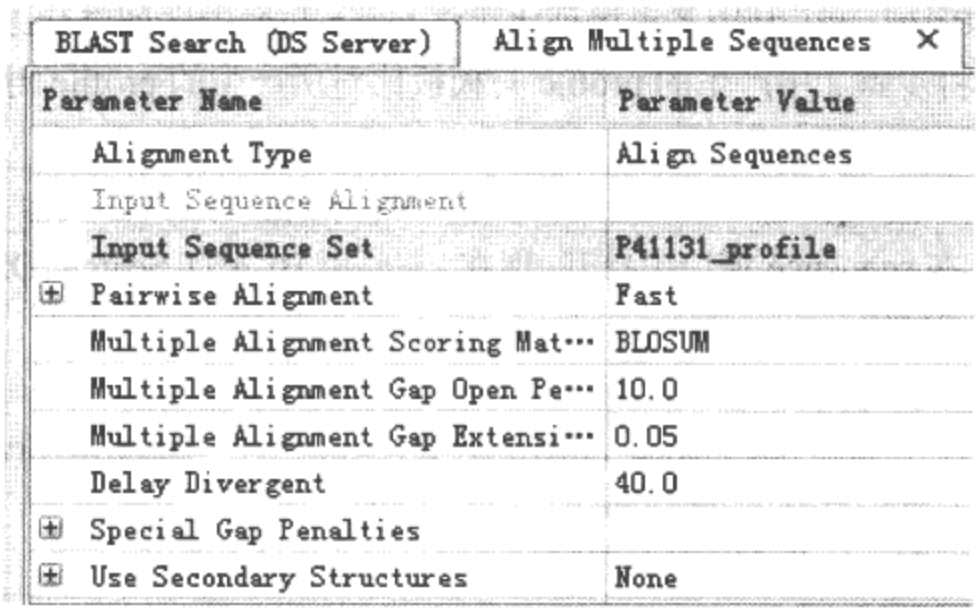


图 3-20　序列比对参数

在 Protocols toolbar 中，点击 ▶ 运行，等待计算完成。计算完成后,在 Jobs 中，双击“completed job”。这将打开一个 Html 的窗口（图 3-21），里面为比对的结果文件 Report.htm。

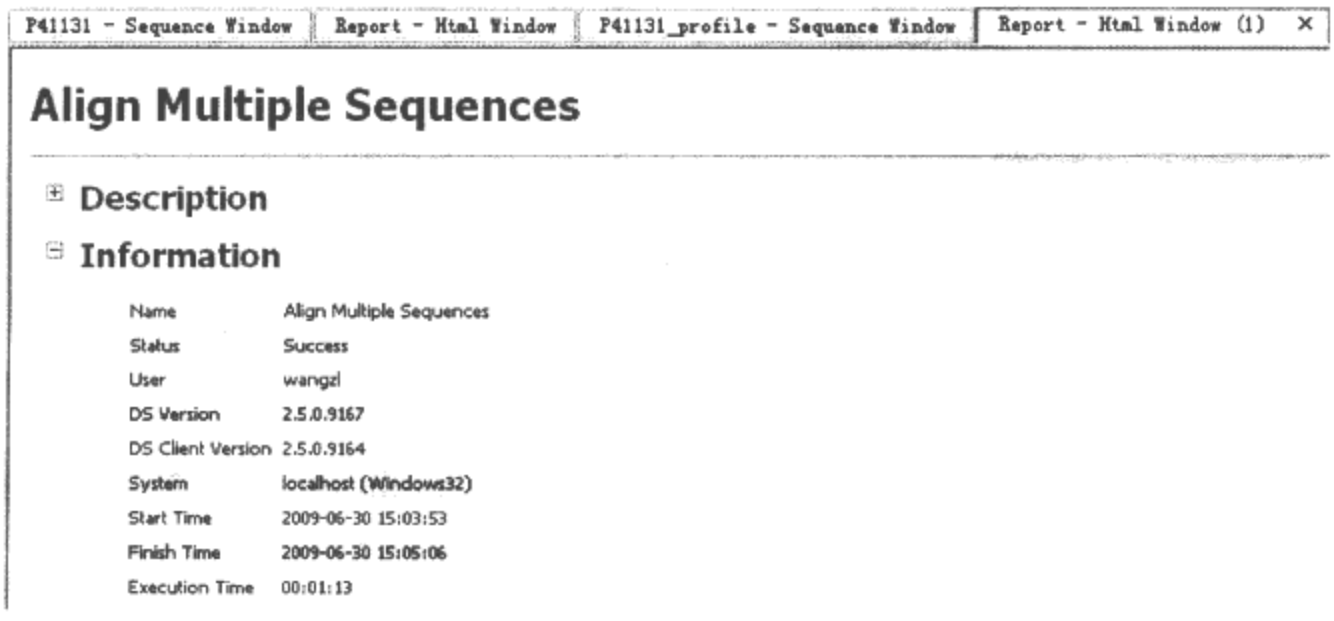


图 3-21　序列比对结果报告

5．将本轮序列比对的 profile 与之前模板结构比对的 profile 进行比对

在 Html 窗口中，Output Files 部分，点击 P41131_profile.bsml 打开序列比对结果。在 Jobs 中，双击之前已经完成的 Align Structure (MODELER) 计算打开 Report.htm 文件。在 Html 窗口，Output Files 部分，点击 View Results 打开序列比对序列比对结果和叠合后的结构。在 Protocols 中，展开 Sequence Analysis 文件，双击 Align Multiple Sequences protocol，点击 Alignment Type 选择 Align Two Profiles，点击 Input Sequence Alignment 选择 P41131_profile（图 3-22），点击 Input Sequence Set parameter 选择 Sequence。

BLAST Search (DS Server) | Align Multiple Sequences ×

Parameter Name	Parameter Value
Alignment Type	Align Two Profiles
Input Sequence Alignment	P41131_profile
Input Sequence Set	Sequence
⊞ Pairwise Alignment	Fast
Multiple Alignment Scoring Mat…	BLOSUM
Multiple Alignment Gap Open Pe…	10.0
Multiple Alignment Gap Extensi…	0.05
Delay Divergent	40.0
⊞ Special Gap Penalties	
⊞ Use Secondary Structures	None

图 3-22　profile-profile 比对

计算完成后，在 Jobs 中，双击 completed job，打开 Report.htm 文件。

在 Html 中，Output Files 部分，双击 P41131_profile-Sequence.bsml 打开序列比对结果。将序列比对结果 P41131_profile 与模板结构比对结果 Sequence 进行 profile-profile 比对。

6．查看 profile-profile 比对结果

计算完成后，在 Jobs 中，双击 completed job，打开 Report.htm 文件。在 Html 中，Output Files 部分，双击 P41131_profile-Sequence.bsml 打开序列比对结果（图 3-23）。

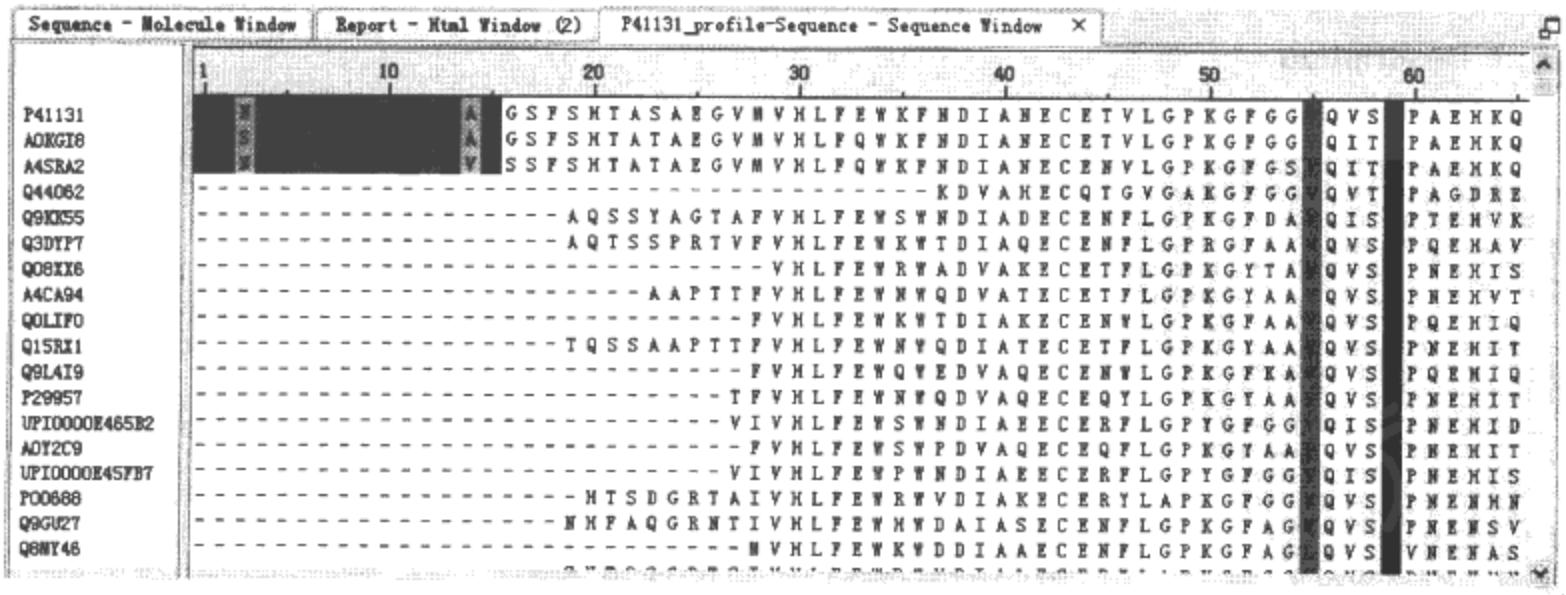


图 3-23　profile-profile 比对结果

四、使用 MODELER 构建目标序列的 3D 模型

在本小节，将使用上小节产生的比对结果构建目标序列 P41131 的 3D 模型。

在 Protocols 中，展开 Protein Modeling 文件， 双击 Build Homology Models，在 Parameters 中，点击 Input Sequence Alignment ，选择 P41131_profile-Sequence。展开 Input Sequence Alignment，点击 Input Model Sequence，选择 P41131，点击 Input Template Structures，选择 4 个模板 1G94A、3DHPA、1HX0A、1JAEA，点击 Optimization Level（图 3-24），选择 low（将“Optimization Level”改为 Low，可加快计算速度，但产生的模型的精度下降）。

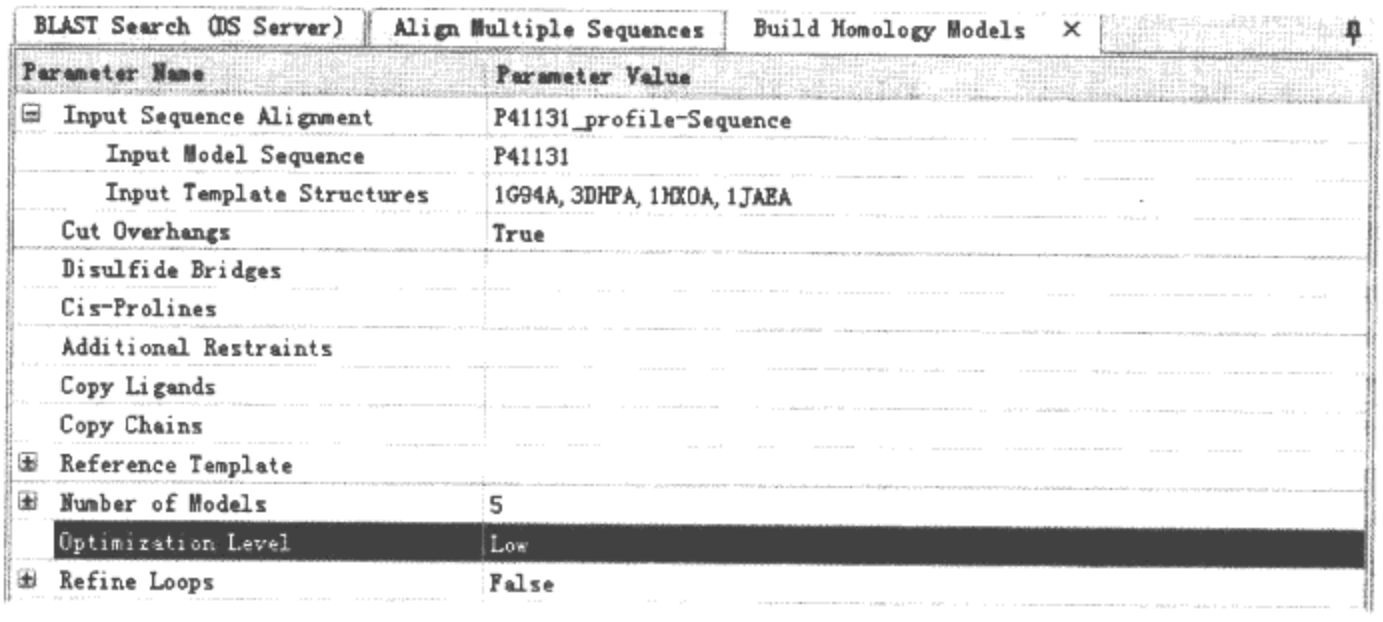


图 3-24　构建模型参数

在 Protocols toolbar 中，点击 ▶ 运行，等待计算完成。

计算结束后，双击 Jobs 中 completed job 打开 Report.htm 文件。

Report.htm 文件（图 3-25），Output 部分，点击 P41131.B99990002.dsv 打开所建 5 个模型中的第一个模型（图 3-26）。

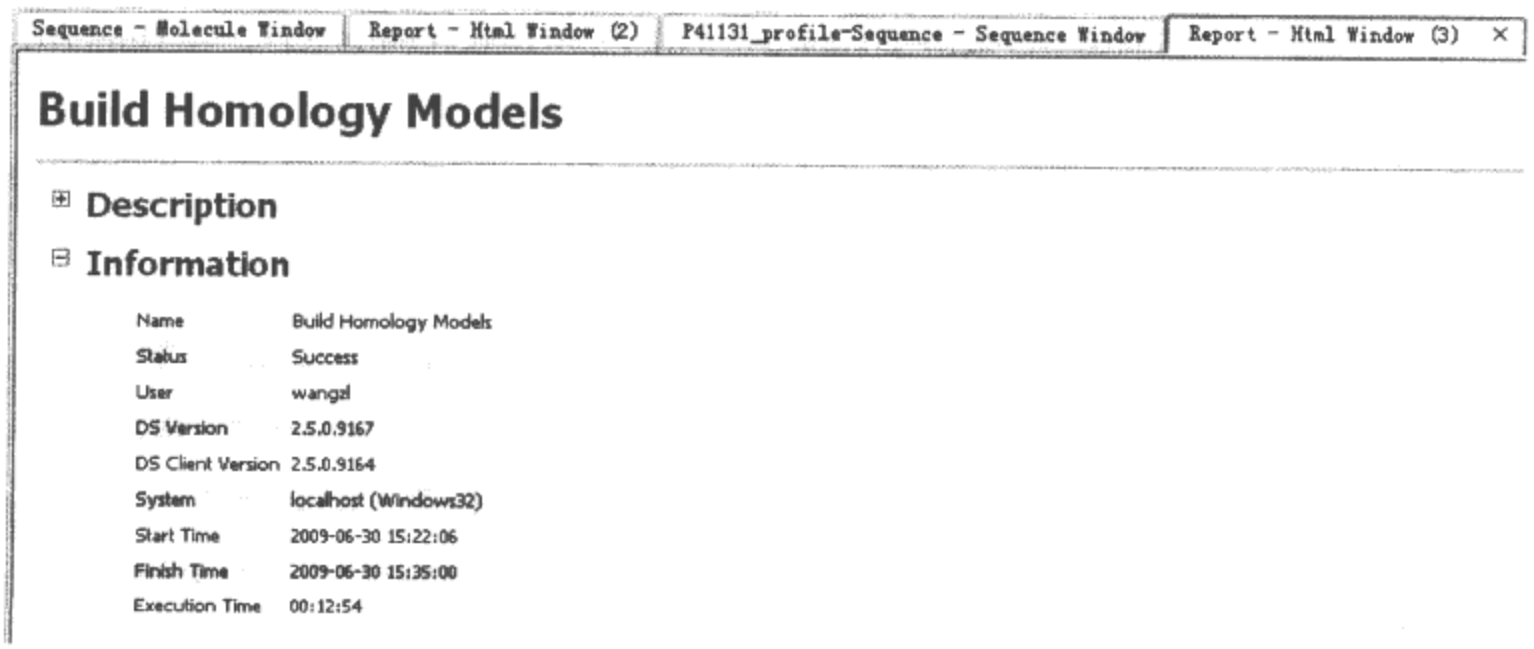


图 3-25　模型构建结果 report 文件

五、模型评估

模型构建完成后，需要对其进行评估。DS 为用户提供了多种模型评估方式。主要包括：PDF（probability density function）Energy、DOPE Score、Ramachandran Plot 和 Profile-3D。

1. 使用 PDF energy 评估和挑选模型

建模过程中，DS MODELER 首先会提取模板（template）的几何特性，然后使用 PDF 函数来定义蛋白结构中诸如键长、键角、二面角等几何特性。接着它会对 PDF 函数施加一定的约束条件，并以此来构建 target 的 3D 结构。所以 PDF 的函数值可以直接反映所构建模型的好坏。一般 PDF Total Energy 越小，表明模型能更好地满足所提取的同源约束条件，模型的可信度越大。

图 3-26　模建的 5 个模型中的第一个模型

2. 查看所有模型的 PDF 值并根据该值选取最优模型

Report.htm 文件窗口，Summary 部分（图 3-27）

该部分有一个表格，表格中给出了模型的 PDF Total Energy、PDF Physical Energy 及 DOPE Score。模型排序时按照 PDF Total Energy 升序排列。

Summary

Modeler Version : 9v4

Models Sorted by PDF Total Energy

Model Name	PDF Total Energy	PDF Physical Energy	DOPE Score
P41131.B99990002	14404.01	1245.35	-43028.44
P41131.B99990005	14512.22	1277.38	-43036.41
P41131.B99990003	14513.43	1250.7	-43354.4
P41131.B99990004	14595.1	1319.73	-43752.91
P41131.B99990001	14840.72	1576.58	-43172.65

图 3-27　评估的 Summary 部分

从图中，可以发现，模型 P41131.B99990002 为最优模型（PDF Total Energy 值最低）。

3. 使用 DOPE Score 评估和挑选模型

DOPE 是一个基于原子统计势能的程序，主要用于模型评估。它的分数可以认为是衡量同一分子不同构象可信度的标准，能够帮助选择预测结构的最优模型。分数越低，模型越可靠。

4. 查看所有模型的 DOPE Score 值并根据该值选取最优模型

Report.htm 文件窗口，Summary 部分（图 3-27）

从图中可以看到，模型 P41131.B99990004 为最优模型。

注意：本实例中，DOPE Score 挑选的最优模型与 PDF Total Energy 挑选的最优模型不一致。所以需要其他评估方法对模型进行评估和挑选。

5. 使用 Ramachandran Plot 评估模型

Ramachandran Plot 用于阐述蛋白质或肽立体结构中肽键内 α 碳原子和羰基碳原子间的键的旋转度（psi）对 α 碳原子和氮原子间的键的旋转度 （phi），主要用来指明蛋白质

或肽类中氨基酸的允许和不允许的构象。

通过对已知晶体结构的统计分析，可确定氨基酸在 Ramachandran Plot 中经常出现的区域。使用该图，用户可以确定结构中每个氨基酸的构象是否正确。本小节只选用所构建的 5 个模型中的 1 个（P41131.B99990002）为例，显示如何调用 Ramachandran Plot。

Chart 菜单，选择 Ramachandran Plot（图 3-28）

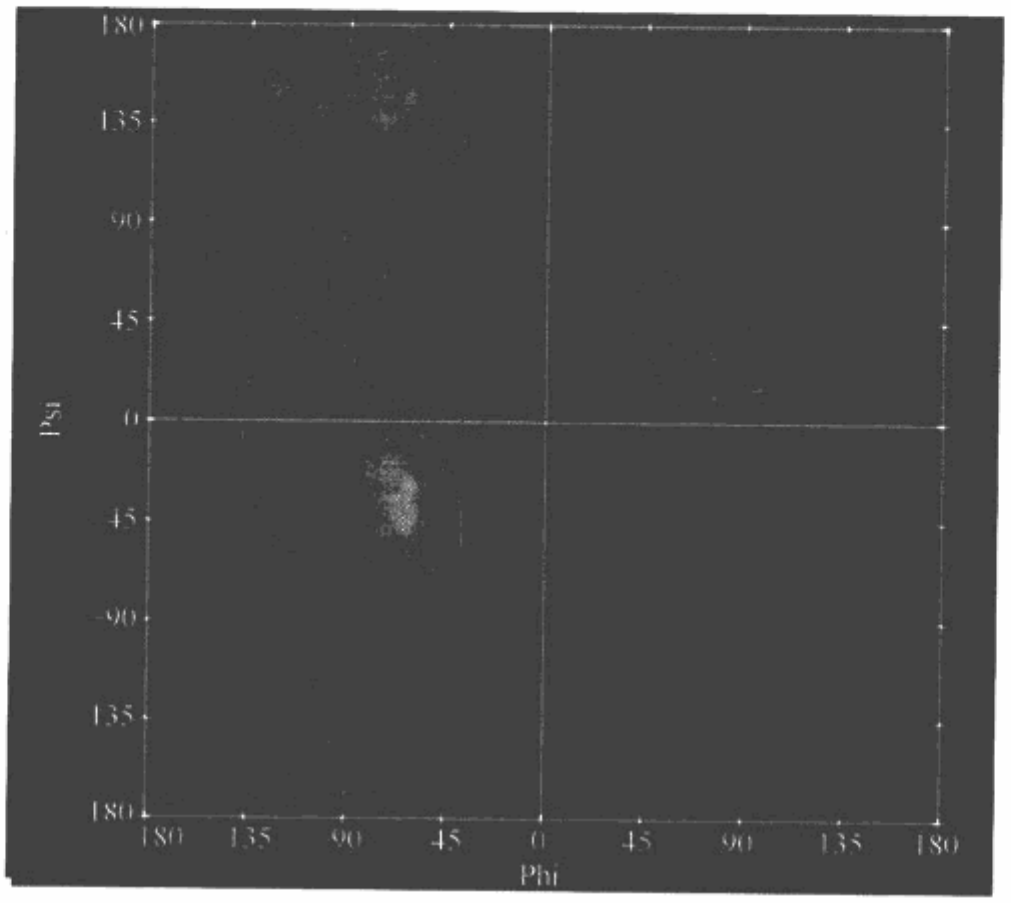


图 3-28　Ramachandran 图

蓝色区域为“最适区”，该区域含有的氨基酸个数越多，结构越可信；紫色区域为“允许区”；其他区域的点（红色）为 psi-phi 构象不合理的氨基酸，可能是建模的错误区，需要优化。

6. 使用 Profile-3D 评估模型

Profile-3D 是加州大学洛杉矶分校（UCLA）的 David Eisenberg 开发的一种基于“穿线”（threading）法的模型评估程序。该方法采用 3D-1D 的打分函数来检测所构建模型与自身氨基酸序列的匹配度关系。分数越高，说明同源模型的可信度越大。本小节只选用所构建的 5 个模型中的 1 个（P41131.B99990002）为例，显示如何使用 Profile-3D 评估模型。

（1）调用待评估模型

点击 Build Homology Models - Html Window 使之激活，该 Html 窗口中，Output Files 部分，点击 P41131.dsv

Output Files

Annotated Model Structure	P41131.B99990001.dsv
Annotated Model Structure	P41131.B99990002.dsv
Annotated Model Structure	P41131.B99990003.dsv
Annotated Model Structure	P41131.B99990004.dsv
Annotated Model Structure	P41131.B99990005.dsv
Best Model Structure Superimposed to Templates	P41131.dsv
Sequence Alignment of Models to Templates	P41131_profile-Sequence.bsml

这将打开一个新的 3D 窗口。里面有最优模型结构（由 MODELER 中的 PDF Total Energy 确定）与所有模板的结构。

该模型与模板已经进行叠合，且以 ribbon trace 的方式显示。模型中，总 PDF 分数不同的区域以不同的颜色和 ribbon 厚度显示。观察结构，可以快速地识别模型中不可信区域（红色）。模板的 ribbon 则根据二级结构上色（图 3-29）。

图 3-29　最优模型与模板的结构

（2）计算模型 Profile-3D 分数

在 Protocols 中，Protein Modeling 文件夹，双击 Verify Protein (Profiles-3D)。在 Parameters Explorer 中，点击 Input Protein Molecules，只选择 P41131:P41131.B99990002，在 Protocols toolbar 中，点击 ▶ 运行，等待计算完成

Jobs 中，双击 Verify Protein (Profiles-3D) job 打开 Report.htm 文件。关闭 P41131-Molecule 窗口，提示保存时选择不保存。在 Html 窗口中，Output Files 部分，点击 P41131.B99990003.dsv

这将打开一个名称为 P4113.B99990002 的新的 Molecule 窗口。里面的结构用不同宽度的 solid ribbon 显示，并根据 Verify Score（得分越高，结构越好）着色。颜色的范围由蓝到白再到红色。蓝色对应高分区，白色为平均分数区域，红色为低分区（图 3-30）。ribbon 的宽度则与 Verify Score 的趋势相反（结构越差，ribbon 越宽）。

（3）查看结果

在 Data Table View 中，点击 Molecule，滚动至 the Verify Expected High Score、Verify Expected Low Score 以及 Verify Score（图 3-31）

Verify Score 越高，模型的质量越好。

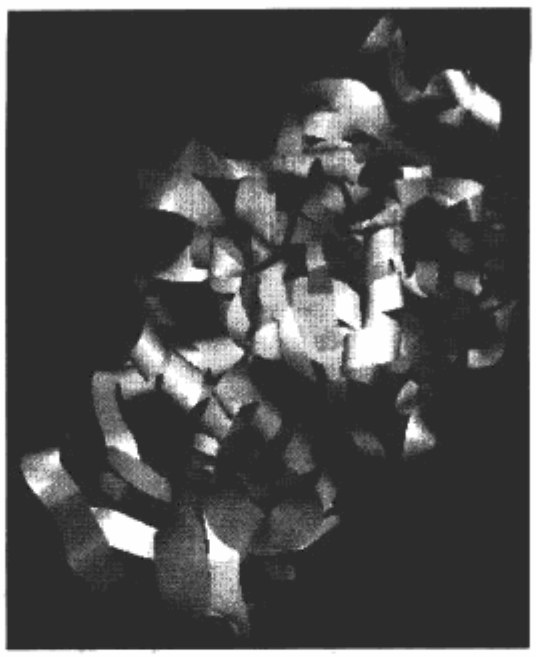


图 3-30　评估模型的 solid ribbon 图

Summary

P41131.B99990002 Verify Score = 171.25

P41131.B99990002 Verify Expected High Score = 195.312

P41131.B99990002 Verify Expected Low Score = 87.8902

图 3-31　评估分数的 Summary

（4）使用图性查看每个氨基酸的分数

点击 AminoAcid，滚动至尾列 Verify Score，点击 Verify Score 的题标选择整列（图 3-32）

	PDF SSVA	PDF STDE	PDF Total	Verify Buried Area	Verify Environment Class	Verify Polar Fraction	Verify Score
1	0	0.6	0.5	39.465	E	0.742	0.32
2	0	1.7	25	28.612	E	0.802	0.32
3	0	2.1	30	65.63	P2	0.74	0.32
4	0	2.3	29	86	P1	0.295	0.32
5	0	4.8	34	110.19	P2	0.675	0.32
6	0	0.88	37	111.076	P1	0.5	0.32
7	0	1.3	23	36.254	E	0.845	0.328
8	0	2.1	28	0	E	0.953	0.226
9	0	4.9	30	20.751	E	0.926	0.428
10	0	2.5	25	134.515	B3	0.659	0.449
11	0	0.37	30	35.189	E	0.651	0.536
12	0	2.5	23	128.931	B1	0.324	0.585
13	0	2.2	21	170.396	B3	0.517	0.458
14	0	2.4	40	130	B1	0.287	0.586

Molecule | Atom | AminoAcid | Group | ProteinSequence | AminoAcidChain | Bond | Assembly

图 3-32　模型中每个氨基酸的分数表

从 menu 中选择 Chart | Line Plot。这将打开一个新的图表（图 3-33），图表中每个氨基酸的 score 都显示出来。注意 C 端和存在两个 loop 的低分区。

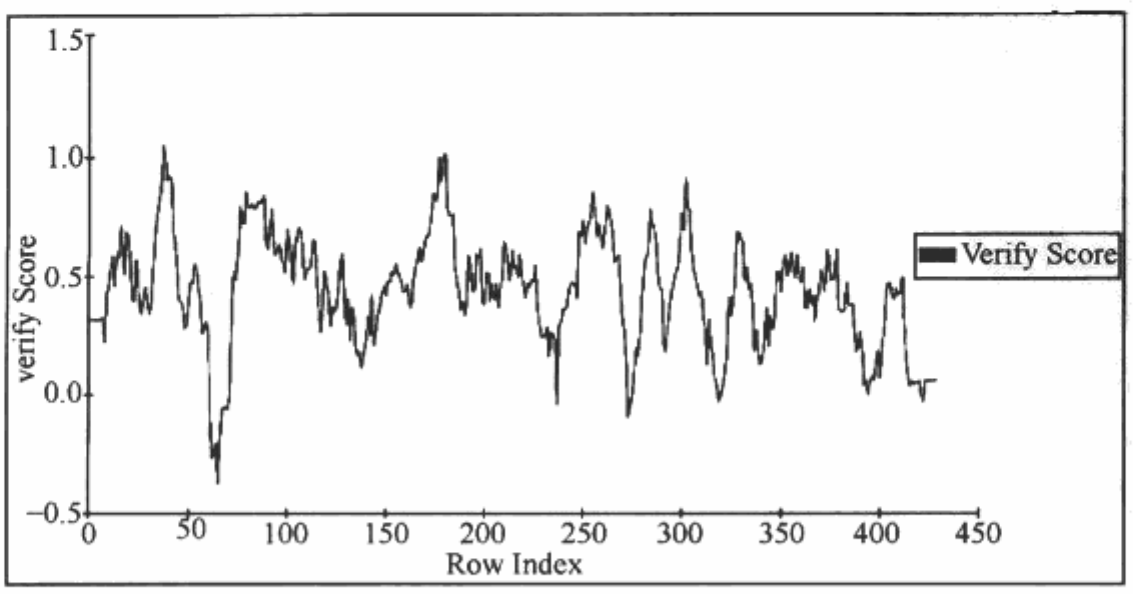


图 3-33　模型的评估分数图

（5）查看低分区的氨基酸

点击 P41131.B99990002-Line Plot 窗口使之激活，点击拖拽 low score peak 周围区域，以选择 Line Plot 中对应的氨基酸

选中的氨基酸在分子窗口中会被标示出来（图 3-34），可以便捷看到低分的比对区域。Profeile-3D 图中选中的低分点对应的氨基酸在分子窗口中直接显示出来（黄色）。

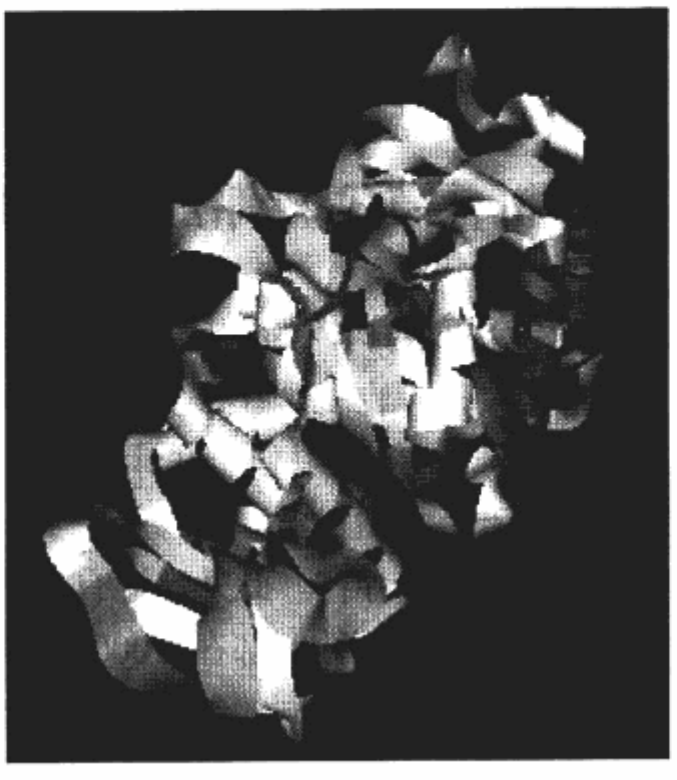


图 3-34　低分区的区域

第四章

利用折叠模式识别法预测蛋白质结构

第一节　折叠模式识别法的基本原理

从理论上讲，一个多肽链可能采取的构象的数目可以是个无限大的数字，但实际上，这些构象中的绝大部分都是能量不利的，也就是在自然界中根本不存在的。许多学者对自然界中可能存在的蛋白质折叠类型（fold）数目作了估计。1992 年 Chothia 估计自然界中折叠类型数量不会超过 1000。1998 年我国科学家王志新院士作了更精确的估计，认为仅有 654 种折叠模式和 1000 种结构超家族存在。此外，还有其他一些专家对自然界中存在的蛋白质折叠模式数量也进行过估计，得到数字都在 1000 至数千。虽然不同科学家对自然界可能存在的折叠类型的具体数目估计有不同的结论，但是大家对这个问题的看法有一些观点是共同的，即自然界中存在的蛋白质折叠类型是个有限的数字，而且折叠类型数量大概不会超过几千种。对已经测定的蛋白质结构进行统计分析可以发现，序列不同的蛋白质可能采用相似的折叠模式。很多未知结构的蛋白可能具有与 PDB 中的分子具有相似的结构，但是目前没有序列的同源信息可以识别这种关系。

折叠模式识别的原理是识别出与一个给定的问询序列相匹配的折叠模式，也就是说不是用序列来预测折叠模式，而是折叠模式要与序列相一致。这涉及几个过程：在一个已知的被称为折叠库的蛋白质结构库中的搜索；为折叠模式打分以及识别适合序列的折叠模式；将问询序列与打分最高的蛋白质进行序列比对。一旦识别到了这样一个模板，过程余下的部分与比较模建的过程相同。

折叠模式识别方法常常是基于序列相似性搜索和结构信息这两方面。例如，3D-PSSM（three-dimensional position-specific scoring matrix，三维位置特异的打分矩阵）方法利用了 PSI-BLAST 算法来发现与问询序列关系较远的序列，并用结构的信息来补充这种搜索，结构方面信息包括二级结构预测和疏水氨基酸具有存在于蛋白质结构内核的趋向。折叠模式识别方法可以用来探测远源关系的序列，但是其结构预测的准确度受到序列比对误差的限制。

蛋白质的折叠模式识别（fold recognition）技术可以确定未知结构的序列对应的折叠模式，或者评价一个序列是否与已知的折叠模式库中的一个结构兼容。这项技术由多家实验室发展而来，所以有很多名称，如反向蛋白质折叠（inverted protein folding）、3D-profiles、Threading 方法、序列-结构兼容性搜索（sequence-structure compatibility search）等。

目前有多种基于蛋白质折叠模式识别预测蛋白质结构的方法。目前利用折叠识别预测蛋白质结构的主要方法大多是从 1991 年 Bowie 等提出的三维剖面（3D-profiles）方法和 1992 年 Jones 等提出的 Threading 方法发展而来的。

第二节　折叠模式识别法预测蛋白质结构的操作实例

Threader 是英国伦敦大学的 David Jones 创立的常用折叠模式程序，曾经在第一届蛋白质结构预测技术的鉴定评估（The Critical Assessment of techniques for protein Structure Prediction，CASP）竞赛中取得优异的表现，被认为是比较成功的折叠模式识别程序。该程序对于学术用户是免费的，可以从互联网上下载。目前程序最新的版本为 Threader3.0。在本章的下面内容中将以幽门螺杆菌 lpp20 蛋白为例，介绍 THREADER 程序的使用方法，包括程序的安装、运行、结果分析等。

一、程序的安装

先登录 http://bioinf.cs.ucl.ac.uk/threader/网站下载相应的文件包括程序、势函数文件、折叠库文件等。并且通过电子邮箱注册学术用户获得 license。

下载后程序后，按照下列步骤在本地机器上安装程序。

% cd /usr/people/yue	进入自己的个人目录
% mkdir threader	创建 threader 子目录
% cd threader	进入 threader 子目录
% gunzip potentials.dat.gz	解压缩势函数文件
% gunzip < threader3-sgi.tar.gz | tar xfo -	安装程序
% mkdir tdb	创建折叠库文件子目录
% cd tdb	进入折叠库文件子目录
% gunzip < ../tdb_oct96.tar.gz | tar xfo -	解压缩折叠库文件
% setenv THREAD_DIR /usr/people/jones/threader	设置环境变量
% vi $THREAD_DIR/threader.lic	编辑 license 文件（将通过邮箱获得的 license 文件粘贴到该文件中，编辑后保存）

二、Threader 程序的运行

本节拟分析的蛋白质是 lpp20，lpp20 为幽门螺杆菌所特有的一种脂蛋白，与幽门螺杆菌的感染、胃癌以及胃癌高发人群都具有密切的关系，而且还具有良好的免疫原性和免疫保护性，因此研究该蛋白质的三维结构具有十分重要的意义。首先以 lpp20 的氨基酸序列为探针，用 Blast 在蛋白质结构数据中搜索同源的参考蛋白，结果在 PDB 中没有发现与 lpp20 具有同源性的参考蛋白，因此预测 lpp20 的结构不能采用同源模建方法。

在利用 Threader 分析蛋白质时，多种序列文件格式都可以用，但最好是 PIR/OWL 格式。下面是本研究实例中所用蛋白质 lpp20 的序列，共有 175 个氨基酸残基：

MKNQVKKILGMSVVAAMVIVGCSHAPKSGISKSNKAYKEATKGAPDWVVGDLE
KVAKYEKYSGVFLGRAEDLITNNDVDYSTNQATAKARANLAANLKSTLQKDLENEKT
RTVDASGKRSISGTDTEKISQLVDKELIASKMLARYVGKDRVFVLVGLDKQIVDKVREE
LGMVKK

在运行程序之前必须确保程序的 license 文件和用于搜索文库的蛋白质链的目录文件在当前目录下，否则将出差错。Threader 3.5 的缺省 license 文件名称为 threader.key，用于

搜索文库的蛋白质链的目录文件在缺省情况下用 psichain.lst 表示。链目录文件也可以给出另外的文件名，但需要在运行程序时说明。

Threader 3.5 程序的控制参数以命令行选项来表示，大部分参数采用缺省值就可以进行搜索。运行 Threader 3.5 最简单的命令行是：

Threader　lpp.seq　lpp -result.out

这将使 lpp.seq 与位于当前目录的 psichain.lst 文件的每一条链去匹配，并将结果输入 lpp-result.out 文件。程序的作业输出可以通过下面的命令直接到一个日志文件中：

Threader　lpp.seq　lpp -result.out > lpp.log

加入 -p 选项将输出计算产生的序列联配：

Threader　–p　lpp.seq　lpp-result.out > lpp. log

这种情况下，计算的能量会总结到 lpp-result.out 中，联配（和其他作业文件）将送到日志文件 lpp.log 中。

如果将–p 改为–pm，即命令为：

Threader　–pm　lpp.seq　lpp-result.out > lpp. Log

则会生成适合于 MODELLER 程序输入文件的联配格式，这样如果能够找到较为适合的折叠模式，就可以很方便地搭建出目标蛋白的三维结构。

三、结果分析

一般情况下，Threader 程序需要运行几个小时。目标序列与折叠库目录文件中的每一条链进行 Threader 之后得到结果文件，下面是输出文件的主要格式：

–33.01 –439.77 –0.20 6.55 –1.52 100.05 –1.11 –1.11 –72.2 174 53.2 99.4 0 12asA0
–64.08 –537.63 0.26 2.05 –0.18 –22.51 0.05 0.05 –42.4 127 84.1 72.6 0 1r1hA0
–76.35 –433.66 0.44 9.44 –2.38 115.28 –1.26 –1.26 –63.5 174 68.2 99.4 0 12ca00
–88.10 –478.43 0.61 –3.60 1.50 –161.14 1.37 1.37 –132.6 168 76.0 96.0 0 12e8H0
–50.32 –505.71 0.06 1.37 0.02 –22.54 0.05 0.05 –2.6 161 83.0 92.0 0 14gsA0
27.65 –301.81 –1.08 –4.58 1.79 –65.37 0.46 0.46 –2.9 144 89.5 82.9 0 1501B0
126.96 –157.46 –2.54 3.80 –0.70 204.09 –2.10 –2.10 –14.1 155 83.8 88.6 0 153100
–16.96 –24.91 –0.43 1.59 –0.04 15.25 –0.31 –0.31 –2.2 102 76.1 58.3 0 155c00
–7.80 –788.48 –0.56 0.99 0.14 12.25 –0.28 –9.99 –17.1 151 48.6 86.3 0 16vpA0
–110.11 –525.51 0.93 –3.73 1.54 –185.82 1.60 1.60 9.6 144 84.2 82.3 0 191400
–78.61 –407.73 0.47 2.57 –0.33 –26.36 0.09 0.09 –67.4 173 59.2 98.9 0 19hcA0
–57.53 –334.56 0.16 –1.59 0.90 –89.84 0.69 0.69 2.6 160 57.1 91.4 0 1a02N0
63.68 –122.51 –1.61 0.48 0.29 73.44 –0.86 –0.86 –68.3 99 97.1 56.6 0 1a02N1
–91.15 –241.56 0.66 –1.60 0.91 –123.69 1.01 1.01 61.5 153 86.0 87.4 0 1a02N2
–129.79 –648.75 1.22 –1.06 0.75 –151.26 1.28 1.28 –21.6 168 82.0 96.0 0 1a04A0
–168.04 –514.73 1.78 –3.60 1.50 –241.23 2.13 2.13 15.4 134 97.1 76.6 0 1a04A2

Threader 程序结果文件主要有 14 列，分别表示不同的参数，每列代表的参数意义如下：

第 1 列是给定的 Threading 的初始两两配对的能量之和，单位是：kcal/mol（1kcal=

4.184kJ）。

第 2 列给出的是模板结构的天然能量（已知结构与其自身序列的能量），如果模板进行 threading 之后获得模型的能量（第 1 列）低于其天然结构的能量（第 2 列），说明是一个错误的匹配。

第 3 列是两两配对之后能量的 Z-分值 (-(energy - mean) / standard_deviation)。

第 4 列为原始溶剂化能量之和。

第 5 列为溶剂化能量的 Z-分值。

第 6 列为经过标准差赋权重之后得到的两两比对及溶剂化的能量的权重加和。

第 7 列是第 6 列的数值的 Z-分值。

第 8 列是第 7 列数值的 Z-分值，其中如果模型与很少的结构匹配或匹配明显错误，则值为–9.99。本列参数是预测结构所依据的最重要的数据。

第 9 列是给出的模型的序列相似性分数（采用的是 BLOSUM62 打分矩阵或折叠库文件中序列的 profile 信息）。如果该分数很高，暗示目标结构和模板结构可能有共同祖先。

第 10 列给出的是目标序列与折叠库中每一条折叠链联配的残基数量。有时候经常需要仔细地寻找在本列中得分最高的模型。

第 11 列是已经联配的结构的百分数。

第 12 列是已经联配的序列的百分数。

第 13 列是局部结构域匹配的起始位置。对于大多数正常的搜索，该列数值一般为 0。

第 14 列是结构在 PDB 数据库中的编码加上对应的链与结构域的标识符。Threader 3 使用的结构域标识符方案与 UCL CATH 数据库中的匹配。新的标识符是 6 个字符，前 4 个字符是对应的 PDB 条目名称，第 5 个字符是折叠模式所处肽链的编号，第 6 个字符是对应的结构域。例如，2tmdA1 指的是 PDB 数据库中 2TMD 条目的 A 链的第一个结构域。如果不需要命名结构域或链（即只有一个结构域或一条链），则缺省的标识符为 0。如 1mbd 00 对应的是 PDB 数据库中的 1MBD 条目，其中有一条链，而且这条肽链没有分结构域（单结构域）。

上述 14 个参数中，对于结构预测结果来说，最有意义的判据是第 8 列参数。表 4-1 是第 8 列参数的不同分值代表的意义。

表 4-1　Z-分值的意义

Z-分值	结 果 说 明
Z > 4.0	很有意义，极有可能是正确的预测
Z > 3.5	有意义，正确预测的机会较大
2.7 < Z < 3.5	处于有意义与无意义的边界，可能是正确的预测
2.0 < Z < 2.7	不理想的分值，只有在其他的证据能够来确认的情况下才有可能是正确的
Z < 2.0	很不理想的分值，在搜索折叠库中可能没有与目标序列匹配的折叠模式
Z = –9.99	与模板折叠模式的序列几乎没有联配的可能

通常可以利用 sort 命令对结果文件进行分类整理，比如，利用命令“sort -n -r +7 lpp.out”可将结果文件按照第 8 列参数的大小排序，下面是对 lpp20 的 Threader 结果文件按照第 8 列参数的大小排序前 10 位的结果。

–278.51 –653.12 3.40　–6.89 2.48　–418.46 3.81　3.81　36.3　157 77.0　89.7　0 2uagA3

–252.03	–576.39	3.01	–7.93	2.79	–413.12	3.76	3.76	16.5	162	73.0	92.6	0	1ryp10
–293.92	–285.43	3.63	–4.69	1.83	–389.24	3.54	3.54	51.2	130	94.9	74.9	0	1ovqA0
–290.57	–710.86	3.58	–3.87	1.58	–369.15	3.35	3.35	16.5	162	71.7	92.6	0	1a8100
–267.75	–468.81	3.24	–4.60	1.80	–361.24	3.27	3.27	46.6	134	85.9	76.6	0	1eejA1
–287.93	–526.17	3.54	–3.55	1.49	–360.06	3.26	3.26	20.0	139	91.4	79.4	0	1squA0
–206.29	–763.79	2.34	–7.38	2.62	–356.03	3.22	3.22	64.5	157	69.5	89.7	0	2tpsA0
–298.72	–465.72	3.70	–2.70	1.24	–353.63	3.20	3.20	81.2	132	86.8	75.4	0	1byrA0
–235.84	–632.97	2.78	–5.70	2.13	–351.64	3.18	3.18	39.6	147	84.5	84.0	0	1g5qA0
–274.58	–655.14	3.34	–3.70	1.53	–349.78	3.16	3.16	–0.4	172	73.5	98.3	0	1e8cA3

从上述结果来看，没有分值在 4 以上的，也就是说没有找到有显著意义的折叠模式，但是有三个折叠模式的分值大于 3.5，分别是 2uagA3、1ryp10、1ovqA0，这也是比较有意义的结果，说明 lpp20 蛋白可能采取三个折叠模式中的某一个。

四、过滤假阳性结果

Threader 预测中可能会出现假阳性结果，除了根据生物学知识过滤假阳性预测之外，Threader 本身也提供了过滤假阳性结果的方法，即“洗牌” 检验（Shuffling Tests）。在序列分析中，经常采用比较一个真实序列和一个随机序列的匹配分数，随机序列一般与真实的序列具有相同的长度和氨基酸组成。通常的情况是通过对原来序列进行“洗牌”来得到随机序列。通过洗牌实验计算的 Z-分值，就可以识别假阳性，突出真的阳性。

在本节中，对于 Threader 分析得到联合 Z-分值大于 2.7 的结果（共 27 个结果）进行了洗牌检验，洗牌检验可以通过-r 命令来实现，-r 后面的数字表示洗牌次数；如果洗牌 50 次，可以用-r50 来实现：Threader –pm -r50 lpp.seq lpp-resultr50.out > lppr50. log

为了节省计算时间，洗牌时可以只对有可能的折叠模式进行重新洗牌计算。

对第一次预测结果中 Z-分值在 2.7 以上的折叠模式进行了洗牌检验，前 10 个结果如下：

–278.515	–257.251	2.89	–6.893	–6.721	1.81	3.29	77.0	89.7	0 2uagA3
–252.028	–178.757	2.86	–7.934	–4.059	3.84	6.14	73.0	92.6	0 1ryp10
–293.920	–225.608	3.56	–4.695	–8.388	1.61	2.91	94.9	74.9	0 1ovqA0
–290.572	–251.088	2.83	–3.870	–9.646	1.13	2.03	71.7	92.6	0 1a8100
–267.755	–244.476	2.58	–4.604	–5.933	1.38	2.42	85.9	76.6	0 1eejA1
–287.932	–187.023	3.68	–3.553	–7.429	1.48	2.67	91.4	79.4	0 1squA0
–206.295	–278.876	1.07	–7.375	–5.809	2.60	1.93	69.5	89.7	0 2tpsA0
–298.723	–326.629	2.19	–2.705	–6.316	0.95	1.70	86.8	75.4	0 1byrA0
–235.837	–253.595	1.91	–5.704	–6.278	1.70	2.64	84.5	84.0	0 1g5qA0
–274.582	–221.603	3.32	–3.704	–6.759	1.23	2.21	73.5	98.3	0 1e8cA3

在洗牌实验结果文件中，第 7 列是联合洗牌 Z-分值（combined shuffled Z-score），是结果中最重要的判断参数，从洗牌检验结果来看，1ryp10 极有可能是 lpp20 蛋白对应的折叠模式。根据 threader 预测得到的目标序列与折叠模式序列的联配结果，利用结构搭建程序如 modeller，就可以搭建出该蛋白的三维结构，进行下一步的分析。

第五章

蛋白质分子对接技术

第一节　蛋白质分子对接概述

近些年来，由于 X 射线晶体结构衍射方法及 NMR 方法的进步，越来越多的蛋白质和核酸的 3D 结构被解析出来。在很多情况下，蛋白质与小分子的相互作用是非常重要的，比如很多酶的底物或抑制剂都是小分子，一些重要蛋白质的拮抗剂也是小分子。从应用的角度来看，可以充分利用一个生物大分子药物靶标的结构和化学特性的信息，采用基于结构的药物设计（structure-based drug design，SBDD）的技术来开发小分子药物（基于结构的药物设计是一种利用蛋白质结构信息来开发和优化小分子抑制剂的迭代的方法）。

获得蛋白靶标与天然底物、天然产物抑制剂、从文库或从头设计的支架得到的化合物的空间结构，就可以精确地确定关键的分子间相互作用。通过了解上述复合物的三维结构，药物化学家和分子模拟专家可以整理出对分子进行理性修饰的策略，以提高化合物的亲和性和选择性，以及改善化合物的药代动力学。然而，生物大分子与其底物的复合物结构却很难得到，这就使得对生物大分子与其相应底物的结合模式的预测显得尤为重要。目前，生物大分子与其底物的结合模式的确定有时候可以依靠实验手段，比如光偶联法、突变法等。例如，2008 年，刘明等人依靠此方法，成功地测定出了多种香豆素（coumarin）类化合物与 HIV-1 整合酶（integrase, IN）的大致结合模式。虽然实验方法能够比较有效地测定生物大分子与相应底物的结合模式，但是由于其耗时、费钱，往往令许多实力较弱的实验室望而却步。此外，更为重要的是，这些方法不能够直接反映出生物大分子与其底物的具体结合模式，而只能给出一个较模糊的信息。有鉴于此，就需要有一种快捷、简便且准确的方法来对生物大分子与其底物之间的相互作用进行预测。分子对接正是这样一种方法。

分子对接计算是把配体分子放在受体活性位点的位置，然后按照几何互补、能量互补以及化学环境互补的原则来实时评价配体与受体相互作用好坏，并找到两个分子之间最佳的结合模式。分子对接是从整体上考虑配体与受体结合的效果，能比较好地避免其他方法中容易出现的局部作用较好而整体结合欠佳的情况。

分子对接在药物设计中具有十分重要的意义。在药物分子产生药效反应的过程中，药物分子与靶酶相互结合，首先就需要两个分子充分接近，采取合适的取向，使两者在必要的部位相互契合，发生相互作用，继而通过适当的构象调整，得到一个稳定的复合物构象。在药物设计中，分子对接方法主要用来从小分子数据库中搜寻与受体生物大分子有较好亲和力的小分子，进行药理测试，从中发现新的先导化合物。

第二节　利用 LibDock 进行分子对接的基本操作方法及实例

DS LibDock 是一个快速的分子对接工具，适用于对大规模数据库进行快速精确的虚

拟筛选。LibDock 根据小分子构象与受体相互作用热区（hotspot）匹配的原理将这些构象刚性对接到受体的结合口袋当中，其最大的优势在于速度快，可以并行运算，适合于进行大规模虚拟筛选，如与 DS Catalyst Conformation 中的 CAESAR 等快速生成构象的方法结合使用更佳。本实例中，一组配体分子将被对接到胸苷激酶（thymidine kinase）中。

一、准备分子对接体系，执行分子对接计算

1．定义蛋白为受体分子

在 Files 中找到文件 1kim.pdb，鼠标双击，该蛋白将在一个三维窗口中出现（图 5-1）。

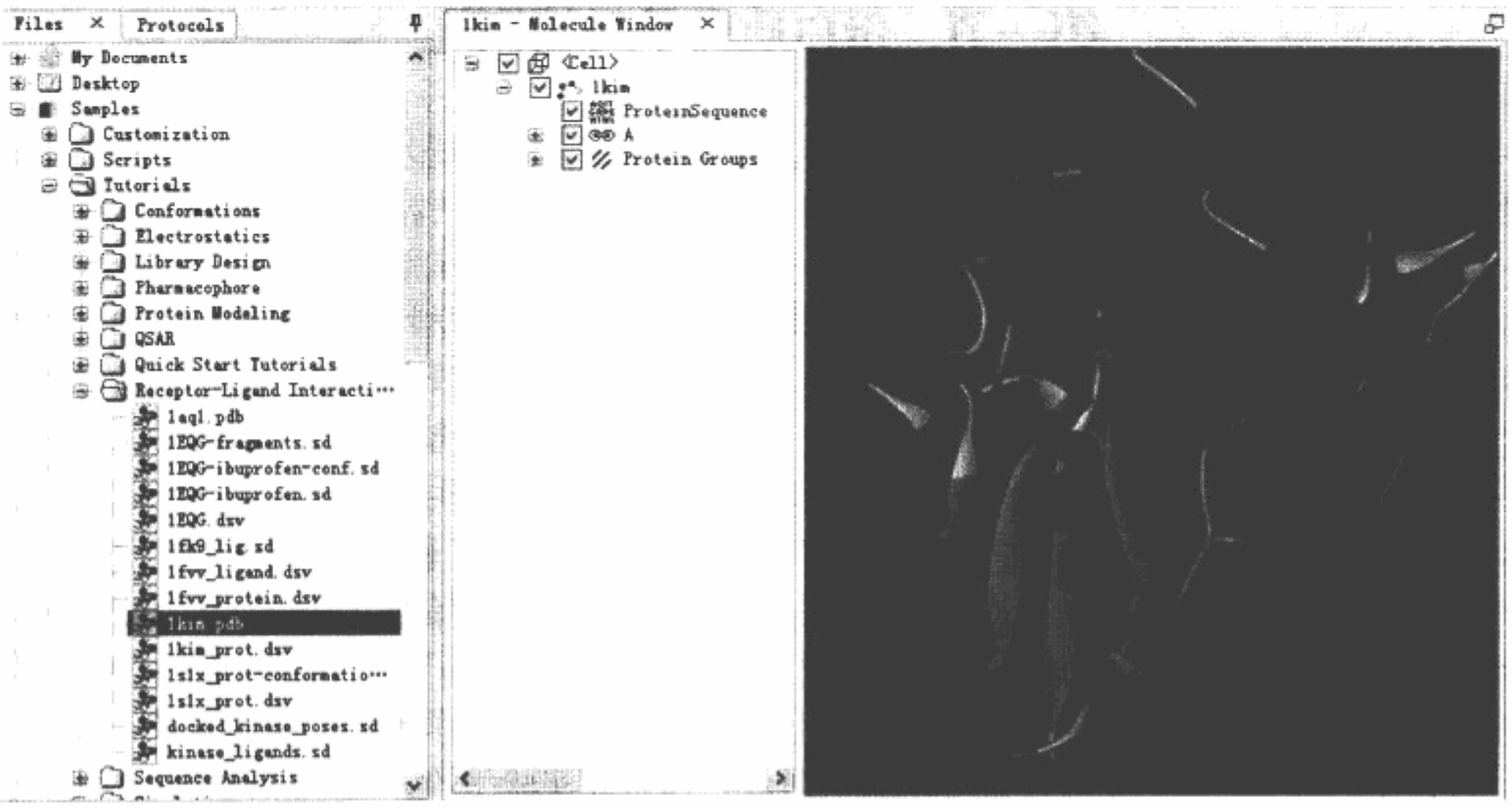


图 5-1　打开蛋白质三维结构

点击选择下拉菜单 Chemistry | Hydrogens | Add Polar 给蛋白添加极性氢。

在系统视图中，展开<Cell>，点击选择 1kim。然后在 Tools 中从下拉列表中选择 Receptor-Ligand Interactions，打开 Receptor-Ligand Interactions 工具面板，在 Define and Edit Binding Site 工具面板下的 Define 工具组中点击 Define Selected Molecule as Receptor 将前面选择的蛋白分子 1kim 定义为受体分子，供下一步使用（图 5-2）。

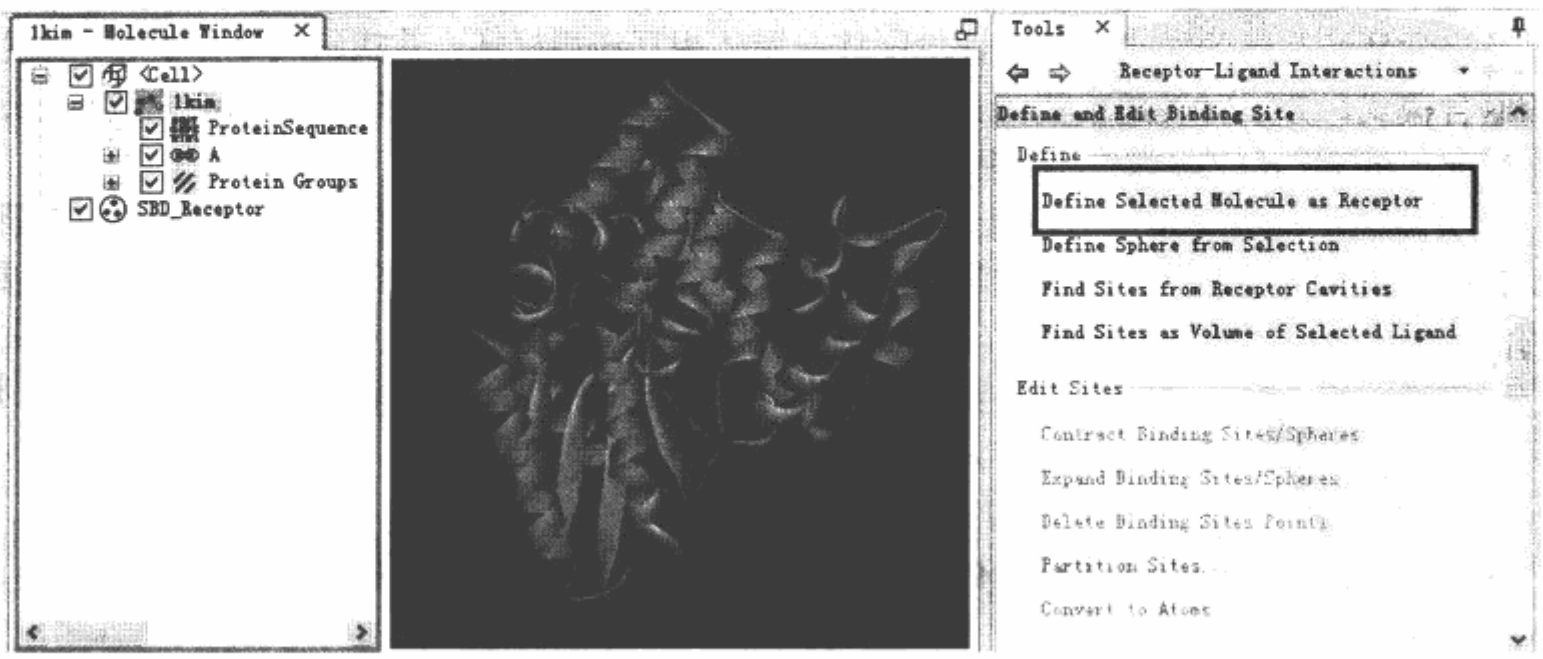


图 5-2　定义蛋白为受体

2. 寻找受体中可能的结合部位

在 Define 工具面板中点击 Find Sites from Receptor Cavities 寻找受体中可能的结合部位（图 5-3）。

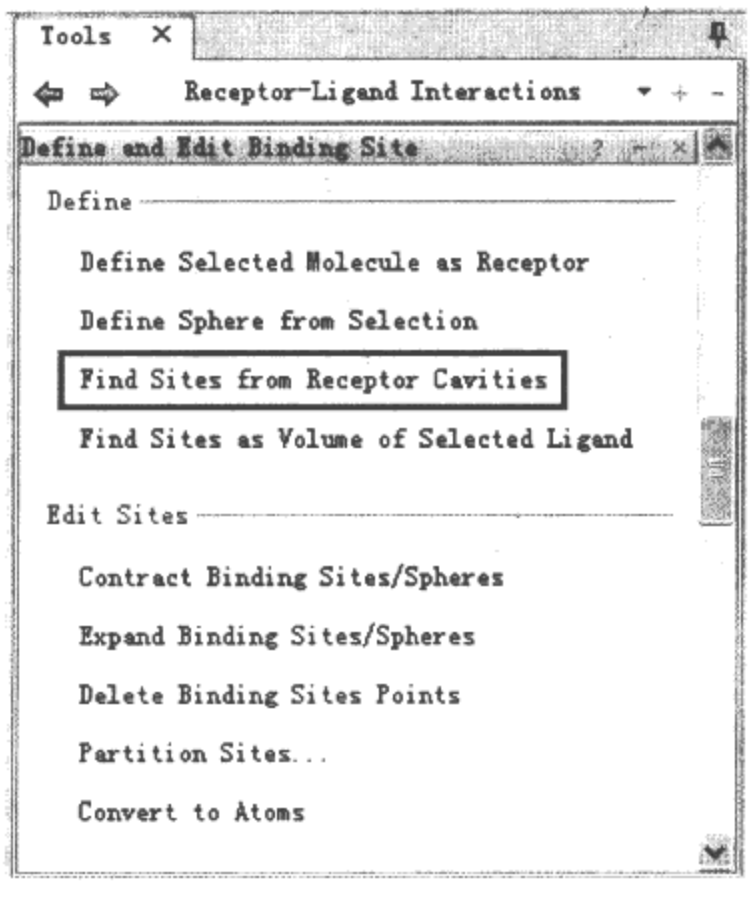


图 5-3　从受体中寻找结合位点

在系统视图中展开 1kim，可以看到识别出 9 个可能的结合位点，最大的可能的结合位点被展示在图形视图中（图 5-4）。

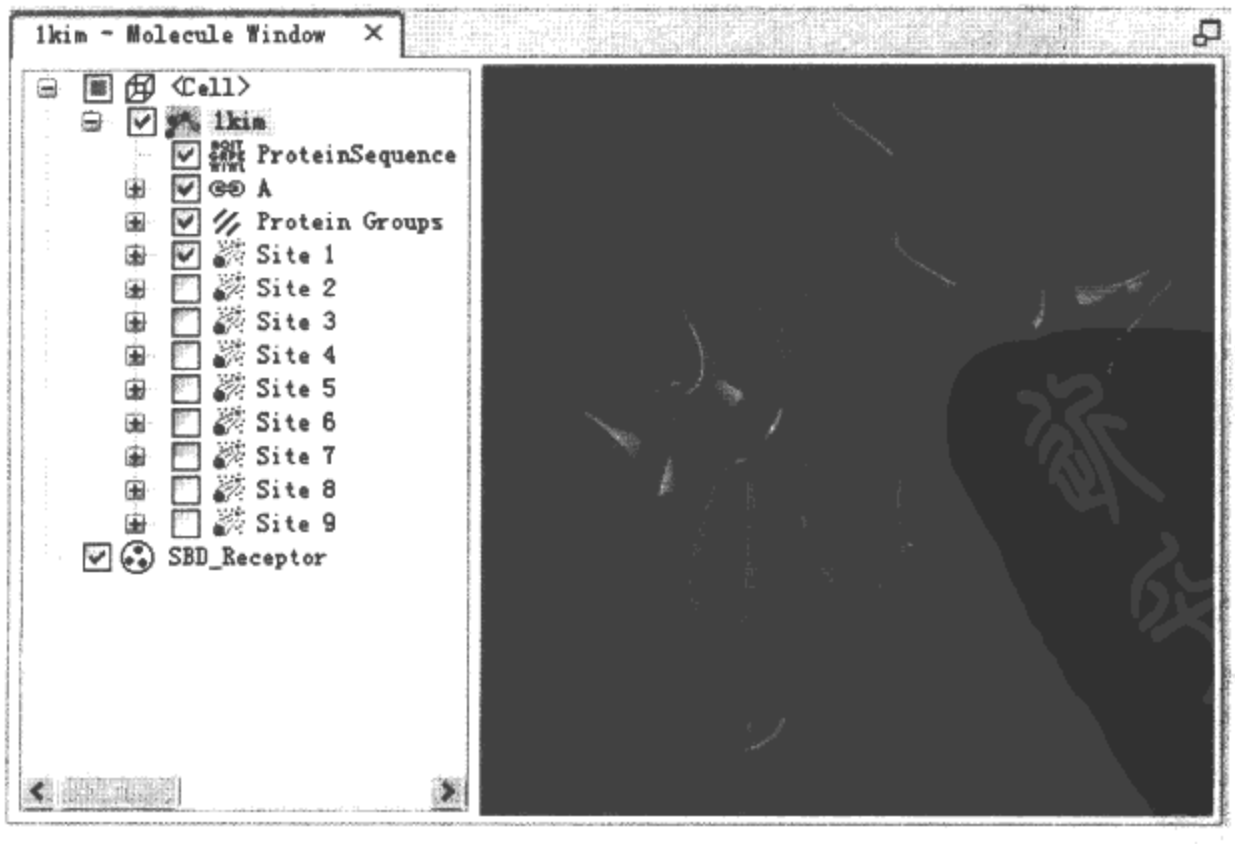


图 5-4　受体结合位点示意图

3. 从结合位点定义 sphere 球

在系统视图中点击选择 Site 1，点击 View 工具栏中的 Fit To Screen 按钮放大结合位点，

然后点击 Define and Edit Binding Site 工具面板下 Define 工具组中的 Define Sphere from Selection，这在结合部位定义出一个红色球，球的名称为 SBD_Site_Sphere（图 5-5）。

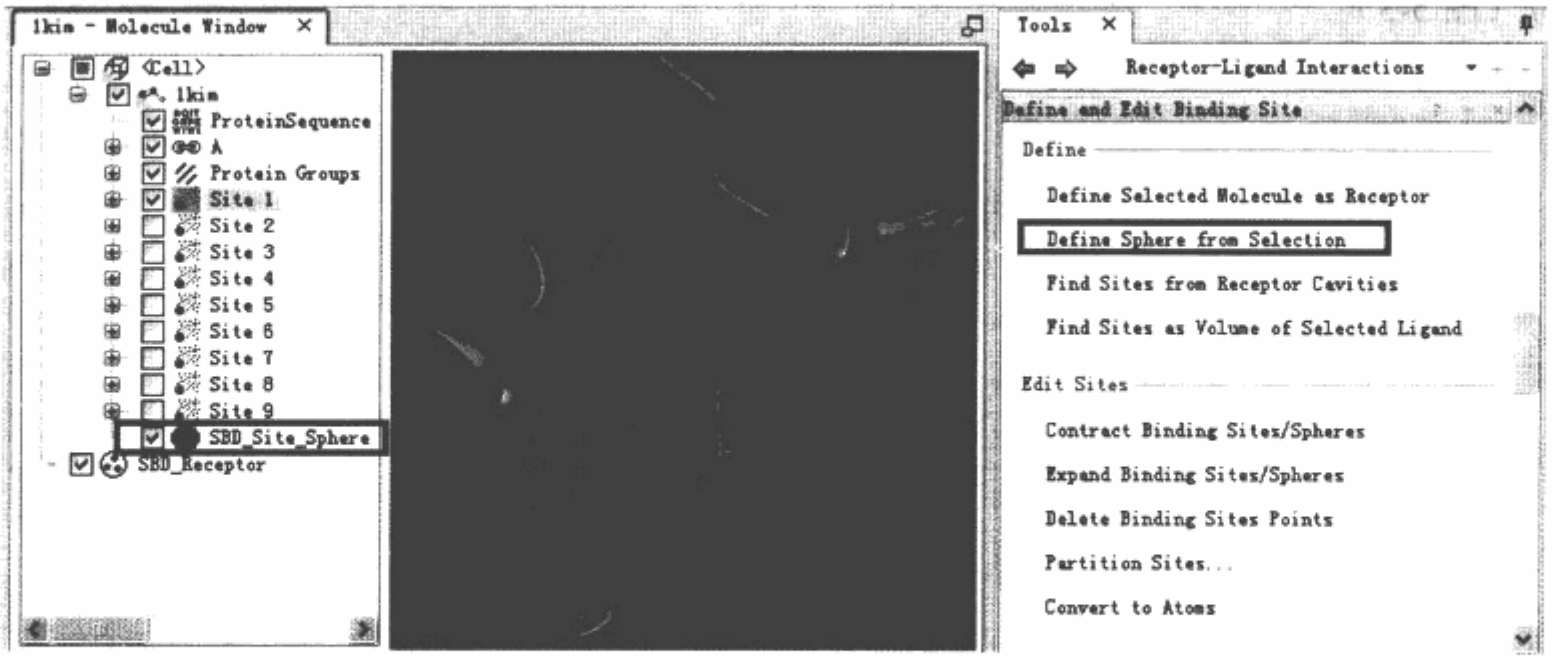


图 5-5　从结合位点定义 Sphere 球

4．修改活性部位球半径

在系统视图中，点击选择 SBD_Site_Sphere，接着单击鼠标右键选择 Attributes of SBD_Site_Sphere...，打开 SphereObjects Attributes 对话框，在对话框中找到半径（Radius）那一行，将值设置为 9，点击 OK 按钮（图 5-6）。

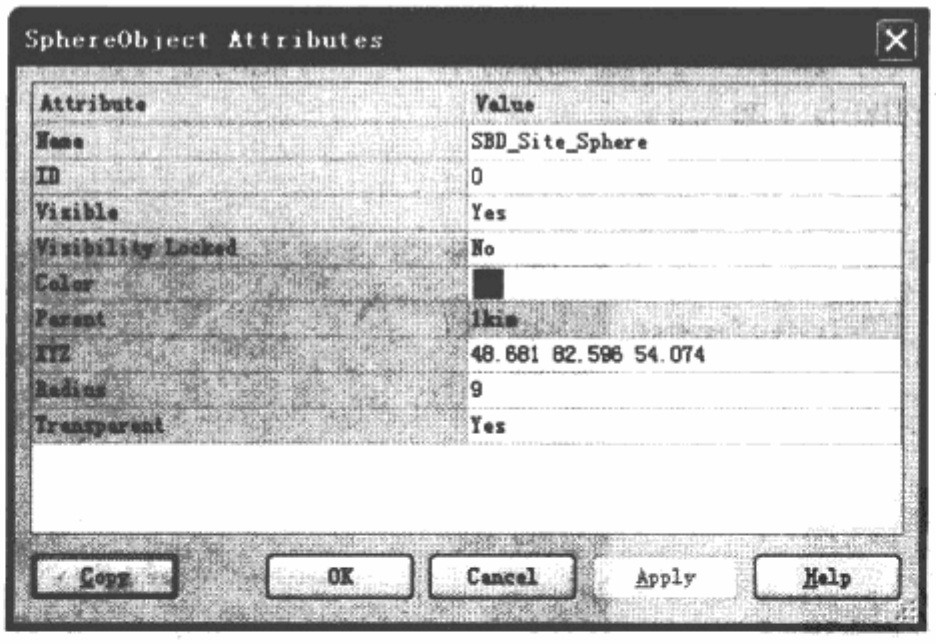


图 5-6　SphereObjects Attributes 对话框

增大半径的目的是为了使得活性部位的大多数点被包含在这个球内，这点可以从视图窗口得到证实。

在系统视图中通过点击 Site 1 前面的框将 Site 1 从三维窗口中隐去，在接着的对接中 Site 1 不会被用到。

按 CTRL+T 隐去窗口中的数据表格（Data Table）。

5．以表格浏览器形式打开配体文件

在文件中找到配体文件 kinase_ligands.sd，双击，配体文件以表格浏览形式打开（图 5-7）。

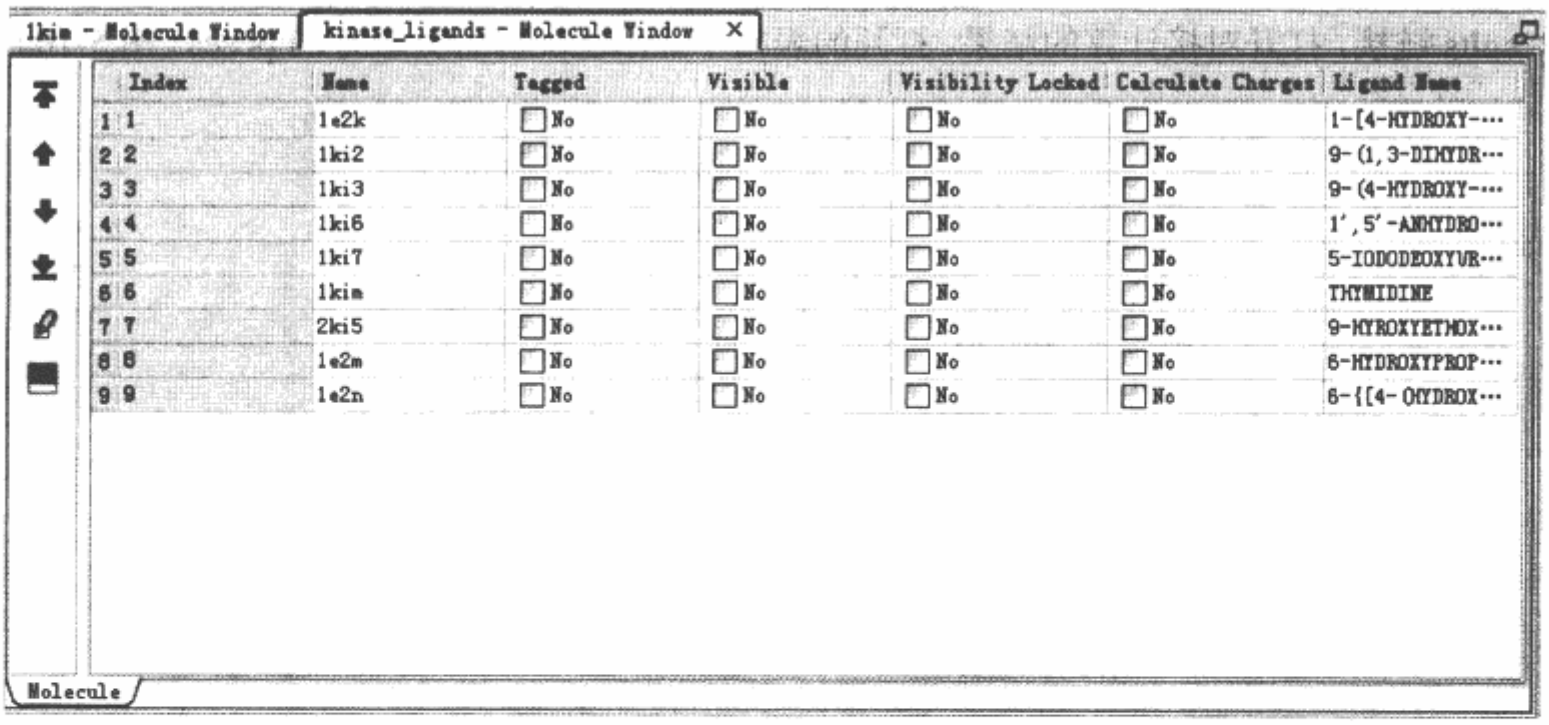


图 5-7　以表格形式浏览待对接配体

6．打开流程，修改参数

在 Protocols 中展开 Receptor-Ligand Interactions 文件夹，双击 Dock Ligands（LibDock）流程，流程对应参数在参数中打开（图 5-8）。

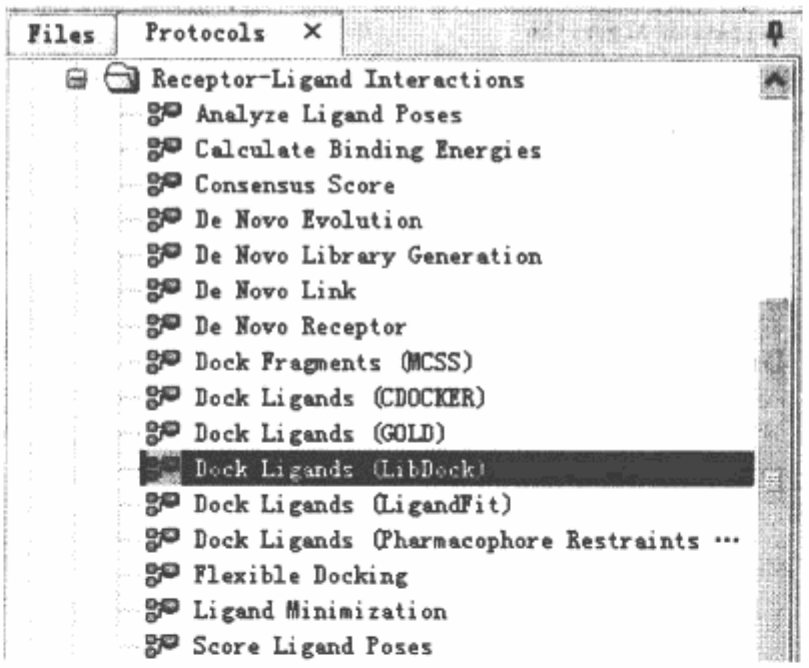


图 5-8　打开 LibDock 对接流程

在参数中，点击 Input Receptor 参数，从下拉列表中选择 1kim:1kim 设置受体蛋白。

点击 Input Ligands 参数，从下拉列表中选择 kinase_ligands:All，指定对接配体。

点击 Input Site Sphere 参数，从下拉列表中选择 sphere 的坐标。

点击 Docking Preferences 参数，从下拉列表中选择 User Specified，这允许你为对接计算改变特定的参数。

展开 Docking Preferences 参数，点击 Max Hits to Save 参数，输入值 10（图 5-9）。

7．运行分子对接计算，查看结果

点击流程工具栏（Protocols toolbar）中的▶按钮运行作业，等待作业完成。

为清晰起见，从菜单栏中选择 Window | Close All 关闭不必要的窗口。待作业完成后，双击作业中刚完成的分子对接作业，打开 Report.htm，点击该文件中 Output 部分的 View

Results 链接，打开对接计算的结果。打开的表格中显示出配体打分最好的 poses（图 5-10）。

Dock Ligands (LibDock)

Parameter Name	Parameter Value
Input Receptor	1kim:1kim
Input Ligands	kinase_ligands:All
Input Site Sphere	48.6806, 82.5963, 54.0744, 9
Number of Hotspots	100
Docking Tolerance	0.25
⊟ Docking Preferences	User Specified
Max Hits to Save	10
Max Number of Hits	100
Minimum LibDockScore	100
Final Score Cutoff	0.5
Max BFGS Steps	50
Rigid Optimization	False
Keep Hydrogens	False
Max Conformation Hits	30
Max Start Conformations	1000
Steric Fraction	0.10
Final Cluster Radius	0.5
Apolar SASA Cutoff	15.0
Polar SASA Cutoff	5.0
Surface Grid Steps	18
⊞ Conformation Method	FAST
⊞ Minimization Algorithm	Do not minimize
⊞ Advanced	
⊞ Parallel Processing	False

图 5-9　LibDock 对接参数设置

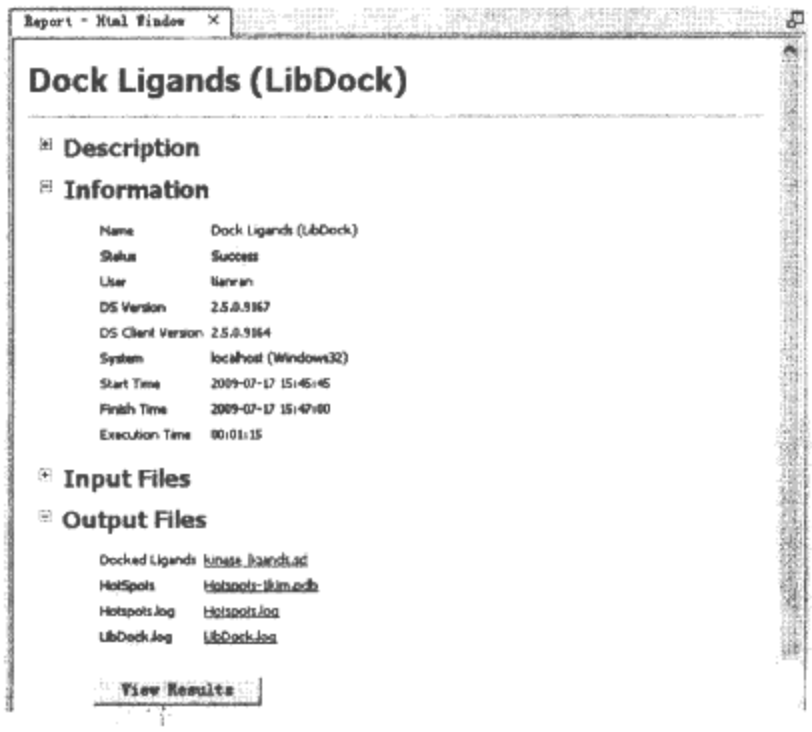


图 5-10　Report.htm 内容

8．浏览对接 poses

点击激活 Molecule-Molecule Window 窗口，确保系统视图和表格浏览器打开，否则按 CTRL+H 和 CTRL+T 打开。按 CTRL＋1 隐藏流程浏览器和作业浏览器。然后在系统视图中删除 Site 文件及 binding site sphere，以便观察结果。

点击表格中 ↑ 键和 ↓ 键将配体在图形窗口中显示出来。在工具栏中点击 按钮，蛋白结合部位被放大。

在 Tools 中从下拉列表中选择 Receptor-Ligand Interactions，打开 Receptor-Ligand Interactions 工具面板，在 Visualize Receptor-ligand Interactions 工具面板下点击 Receptor-Ligand Hydrogen Bonds（图 5-11），在视图窗口中，受体原子与配体对接 poses 间的氢键会通过绿线显示出来（图 5-12）。

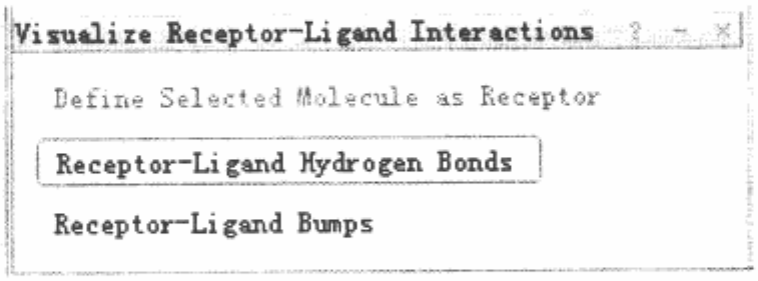


图 5-11　显示蛋白-配体之间氢键

为了更好地观察受体分子与配体对接 poses 间的相互作用，可以对体系进行旋转以获得最佳的观赏角度。在表格中，点击 ↑ 键和 ↓ 键浏览其他的配体对接 poses（图 5-12）。

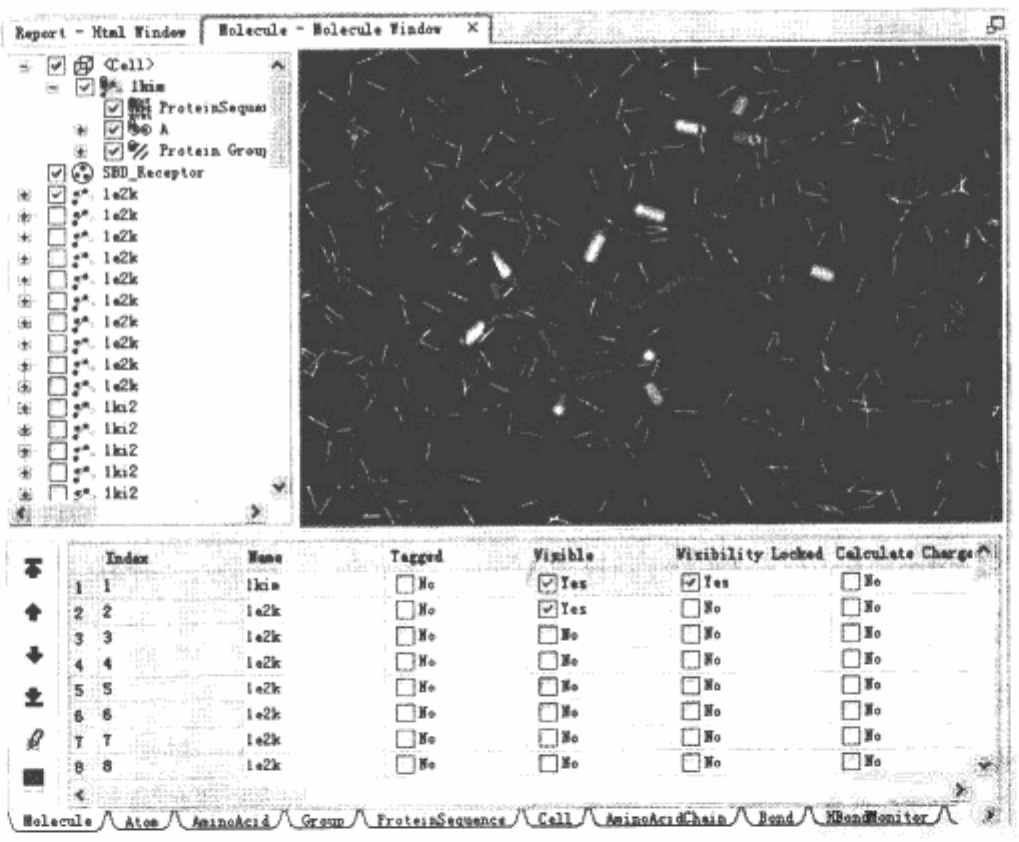


图 5-12　浏览对接结果

二、分析对接结果

使用 Analyze Ligand Poses 流程，计算对接构象的 RMSD 值，受体与对接构象间形成的氢键以及紧密接触（范德华碰撞）情况。

1．打开 Analyze Ligand Poses 流程，修改参数

按 CTRL+1，显示流程（Protocols）和作业（Jobs），打开 Protocols 下的 Receptor-Ligand Interactions 文件夹，双击 Analyze Ligand Poses 流程，相应的参数出现在参数浏览器中。

点击 Input Ligands 参数格，下拉列表中选择 Molecule:All，选中 Molecule 窗口中所有配体构象。

点击 Input Receptor 参数格，从下拉列表中选择 Molecule:1kim。

展开 Input Receptor 参数，点击 Hydrogen Bond 参数格，从下拉列表中选择 True。

展开 Hydrogen Bond 参数，设置 Scope 参数为 Residue 和 Molecule。对接得到的姿态与整个受体之间及受体的每个氨基酸残基形成的氢键数目将都被计算出来。

点击 Contacts 参数格，从下拉列表中选择 True。

展开 Contacts 参数，点击 Type 参数，从下拉列表中选择 All。这一缺省选项计算每个对接 poses 与整个受体及每个对接 poses 与受体中单个氨基酸残基间的极性和非极性接触。点击 Scope 参数格，选择 Residue（图 5-13）。

Analyze Ligand Poses

Parameter Name	Parameter Value
Input Ligands	Molecule:All
Input Receptor	Molecule:1kim
Hydrogen Bonds	True
Scope	Molecule, Residue
Label	HBOND Count
Contacts	True
Type	All
Threshold Definition	VDW Fraction
Scope	Residue
Label	Contacts
RMSD	True
ROC Curve	False
Parallel Processing	False

图 5-13 “Analyze Ligand Poses”流程参数设定

2．运行流程和查看结果

在流程工具栏中，点击▶按钮开始计算。当作业完成时跳出作业完成对话框，关闭该对话框。为简洁方便起见，在菜单栏中选择“Window | Close All”关闭不必要的窗口。

在作业浏览器中，双击已完成的作业，在 DS 的 Html 窗口中打开 Report.htm 文件。在 Html 窗口中“Output Files”部分点击“View Results”链接，分析计算的结果将以表格浏览器的形式打开，其中有些列中包括刚计算得到的性质：

- 对接 poses 与整个蛋白形成的氢键数
- 对接 poses 与受体每个氨基酸残基形成的氢键数（304 列）
- 对接 poses 与整个蛋白间的过近接触
- 对接 poses 与受体每个氨基酸残基间的过近接触（304 列）

点击激活 Molecule-Molecule Window 窗口，按 CTRL+1 隐藏其他浏览器使窗口最大显示。

3．产生氢键数热图

在表格中，点击 HBOND Count 1 1kim:A:MET46 列头选择该列。然后使用滚动条滚动到 HBOND Count 列末，按 SHIFT 键，点击 HBOND Count 304 1kim:A:ALA375 列头，这将选择表格中所有包含氢键数的栅格，其中每一列对应一个氨基酸残基，每一行对应一个 poses。

在菜单栏中，选择 Chart | Heat Map 产生热图，热图中的热点对应着对接 poses（Y 轴）与受体氨基酸残基（X 轴）形成的氢键，比如可以从图中看到 Gln125 与许多对接 poses 形

成氢键（图 5-14）。

通过手动浏览所有的配体对接 poses 来识别这么多的氢键是非常具有挑战性的。热图快速显示氢键图谱，用于洞察那些对配体结合有重要贡献的氨基酸残基。

移动鼠标指向热图中特定的栅格，浏览栅格对应的信息，如残基名称、poses 数和氢键数等。

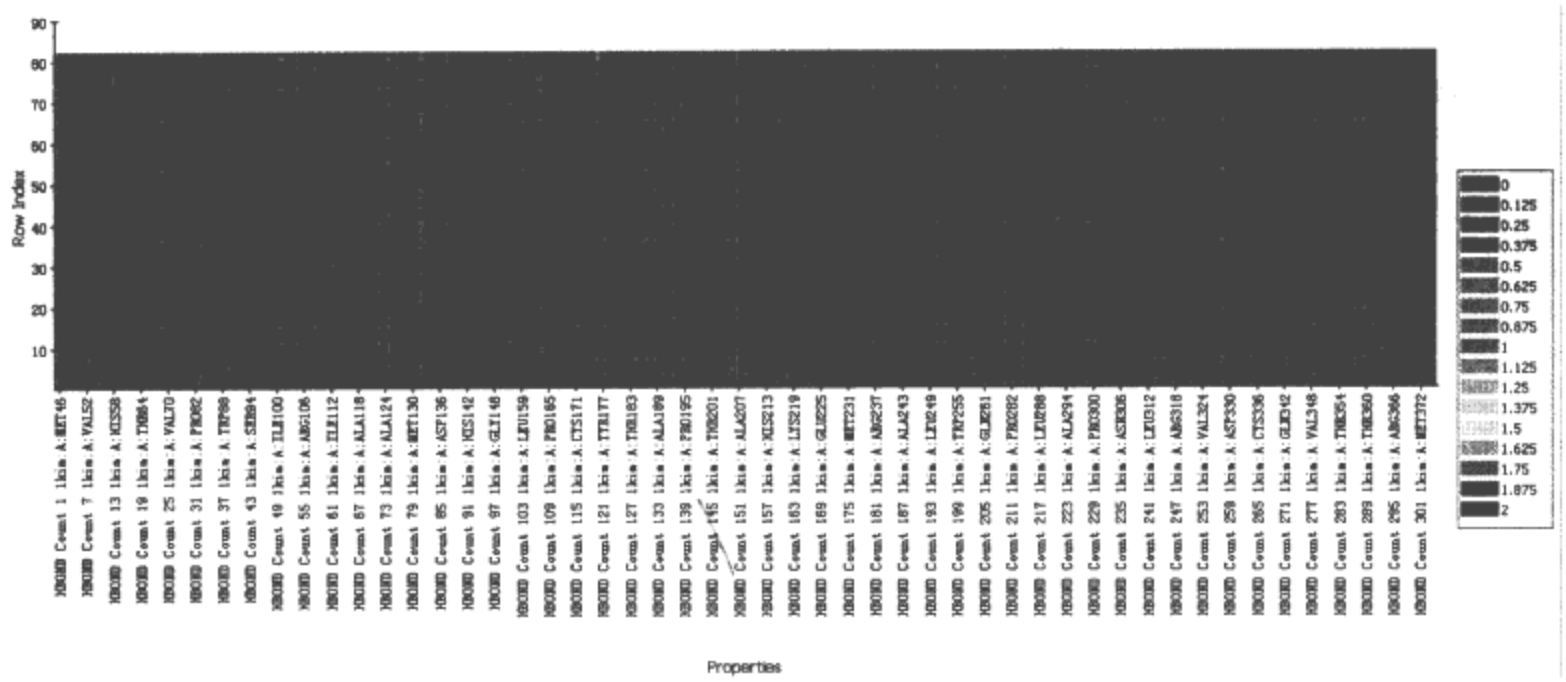


图 5-14　氢键数热图

4．产生过近接触数热图

点击激活 Molecule-Molecule Window 窗口。在表格中，点击 Contacts 1 1kim:A:MET36 列头选择该列。然后使用滚动条滚动到 Contacts 列末，紧按 SHIFT 键，点击 Contacts 304 A:ALA375 列头，这将选择表格浏览器中所有包含过近接触数的栅格，同前面一样，其中每一列对应一个氨基酸残基，每一行对应一个 poses。

在菜单栏中，选择 Chart | Heat Map 产生热图，热图中的热点对应着对接 poses（*Y* 轴）与受体氨基酸残基（*X* 轴）之间的过近接触数目。移动鼠标到热图中的热点，可以看到某个 poses 的过近接触数目，垂直条纹表明某个残基与对接 poses 间的过近接触情况（图 5-15）。

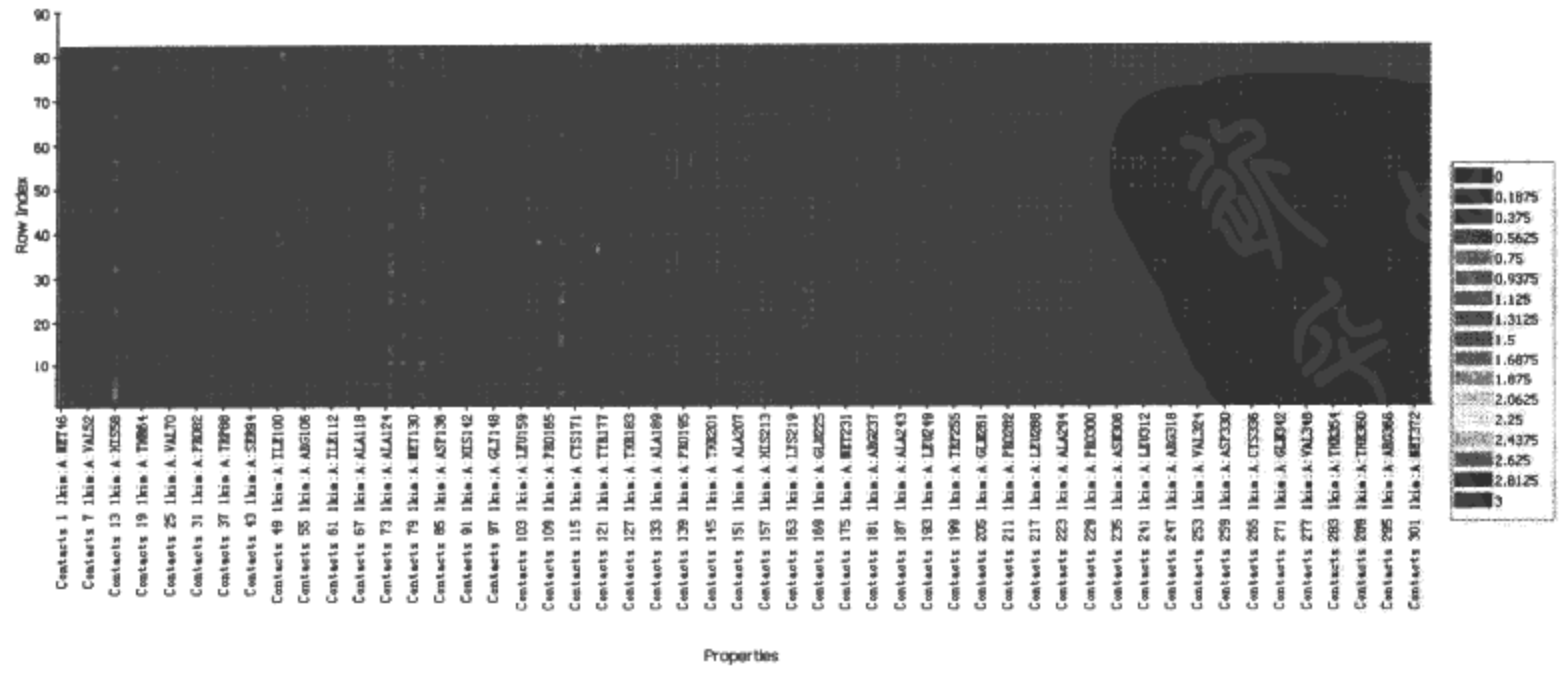


图 5-15　过近接触数热图

图 5-16　过近接触示意图

5．在图形窗口浏览过近接触

点击激活 Molecule-Molecule 窗口。点击表格中 ⬆ 键和 ⬇ 键将配体在图形窗口中显示出来。Tools 中 Receptor-Ligand Interactions 工具面板下，在 Visualize Receptor-ligand Interactions 工具面板下点击 Receptor-ligand Bumps，在视图窗口中，受体原子与配体对接 poses 间的氢键会通过紫色线显示出来(图 5-16)。

点击表格中的其他栅格显示被选 poses 与受体原子之间的过近接触。本实例中不是所有 poses 都会有过近接触存在。

第三节　利用 Flexible Docking 进行受体和配体柔性的分子对接实例

前面介绍的 LigandFit 是一个刚性的对接工具，除了 LigandFit 外，操作系统还提供了一个受体柔性的对接工具，DS Flexible Docking 就是一个受体柔性对接工具（Induced-Fit 对接工具），可以实现配体-受体双柔性对接，模拟配体与受体结合时的诱导-契合效应（induced-fit）。Flexible Docking 模块先对蛋白活性口袋的侧链产生多个构象，然后将配体对接到受体的活性口袋当中，最后对得到的配体-受体复合物结构进行优化。Flexible Docking 最大的优势在于准确，可以精细地研究配体-受体的相互作用信息，适合于作用机理研究。需要 DS CHARMm、DS LibDock、DS Catalyst Conformation、DS Protein Refine 模块支持。

在 Flexible Docking 中，配体将按如下步骤对接到活性部位氨基酸残基有柔性侧链的受体中：①选择受体柔性氨基酸残基，使用 CHARMm 通过改变侧链构象计算出一组结构；②使用 LibDock 将柔性配体对接到每一个受体构象的活性部位；③使用 ChiRotor（CHARMm）在刚性配体存在的情况下优化选定的蛋白侧链；④使用 CDOCKER 优化最后的配体。

在本实例中，你将完成一个晶体重构实验，将来自 PDB ID 为 3ert 中的配体重新对接到受体结构中。

一、准备对接的分子体系，完成分子对接计算

1．在 DS 窗口中打开蛋白和配体文件

在 DS 文件浏览器中，寻找蛋白数据文件 1err_prot.dsv，双击，蛋白将在一个新的三维窗口中打开。接着找到配体数据文件 3ert_ligand.sd，将其拖拽到受体蛋白所在的三维窗口。此蛋白文件已经过前处理，定义好结合位点 Sphere，并且定义了配体 5Å 范围内的氨基酸残基为一个 group，命名为 1err_res1，作为柔性对接的氨基酸残基（图 5-17）。

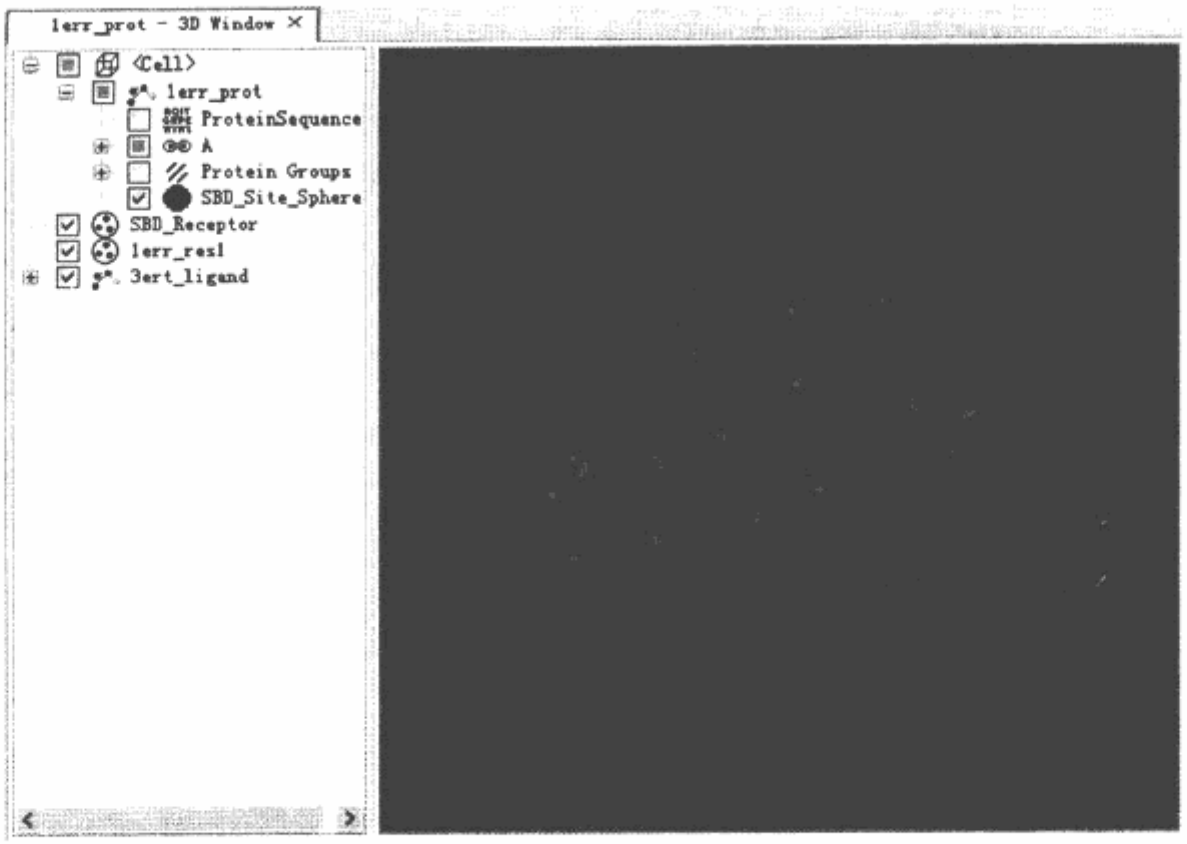


图 5-17　打开蛋白和配体文件

2．对受体及配体分子赋力场参数

在工具（Tools）中，从下拉列表中选择 Receptor-Ligand Interactions，打开 Receptor-Ligand Interactions 工具栏。在 Simulate Structures | Forcefield 工具栏中从下拉列表中选择 CHARMm，点击 Apply Forcefield，给受体及小分子配体赋力场。配体赋好力场后，Tools 中其力场状态将显示为：1err_prot typed with CHARMm（图 5-18）。

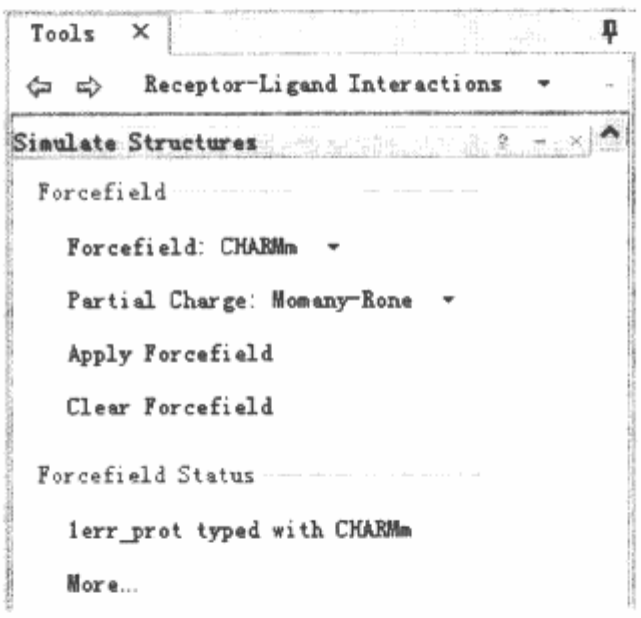


图 5-18　赋予力场参数

3．打开柔性对接流程，修改参数

在 Protocol 中展开 Receptor-Ligand Interactions 文件夹，双击 Flexible Docking 流程，打开柔性对接流程，该流程的参数项在参数中出现。在参数中点击 Input Typed Protein Molecule 参数，然后从下拉菜单中选择 1err_prot:1err_prot。

点击 Input Ligands，从下拉菜单中选择 1err_prot:All。

点击 Input Site Sphere 参数格，从下拉菜单中选择 Sphere 坐标。

点击 Selected Residues 参数格，从下拉菜单中选择 1err_res1。这一操作将选择活性部位的一组氨基酸残基，这些氨基酸残基已经预先被选上且定义为 1err_res1 组。

展开 Generate Protein Conformations 参数组，点击 Maximum Number，输入数值 2。

展开 Generate Ligand Conformations 参数组，点击 Conformation Method 参数，从下拉菜单中选择 best 方法。

展开 Docking 参数组，点击参数 Max Hits to Save 设置为 3（图 5-19）。

4．分子对接计算和结果浏览

在流程工具栏，点击运行按钮▶，等待作业完成。当作业完成时会出现作业完成对话框，关闭作业完成对话框。

Flexible Docking ×

Parameter Name	Parameter Value
Input Typed Protein Molecule	1err_prot:1err_prot
Input Ligands	1err_prot:All
Input Site Sphere	68.026, 34.1162, 74.5856, 9
Selected Residues	1err_res1
⊟ Generate Protein Conformations	True
⊞ Maximum Number	2
Selected Residues	
Stop Protocol	False
Minimum Angle	38
Maximum Number of Residues	8
Energy Range	
Input File	
⊟ Generate Ligand Conformations	True
Conformation Method	BEST
Maximum Conformations	255
Discard Existing Conformat…	True
Energy Threshold	20.0
⊟ Docking	
Number of Hotspots	100
Max Hits to Save	3
Tolerance	0.25
⊞ Preferences	User Specified
RMSD Filter	2.0
⊞ Docking Refinement	
⊞ Advanced Settings	
⊞ Parallel Processing	False

图 5-19　Flexible Docking 参数设置

在开始结果浏览之前，为了简洁起见，在菜单中选择 Window | Close All 关闭所有窗口。

在作业中，双击刚完成的作业，这会在 Html 窗口中打开 Report.htm 文件。在 Html 窗口中的输出文件（Output Files）部分，点击 View Results 链接，打开对接结果。

5．浏览配体对接姿态

点击标签激活结果窗口，为便于浏览，按 CTRL+1 隐藏流程浏览器和作业浏览器。

按 CTRL+H 打开系统视图（Hierarchy View）。

在系统视图中，勾选 SBD_Receptor，蛋白所有氨基酸残基都显示出来。

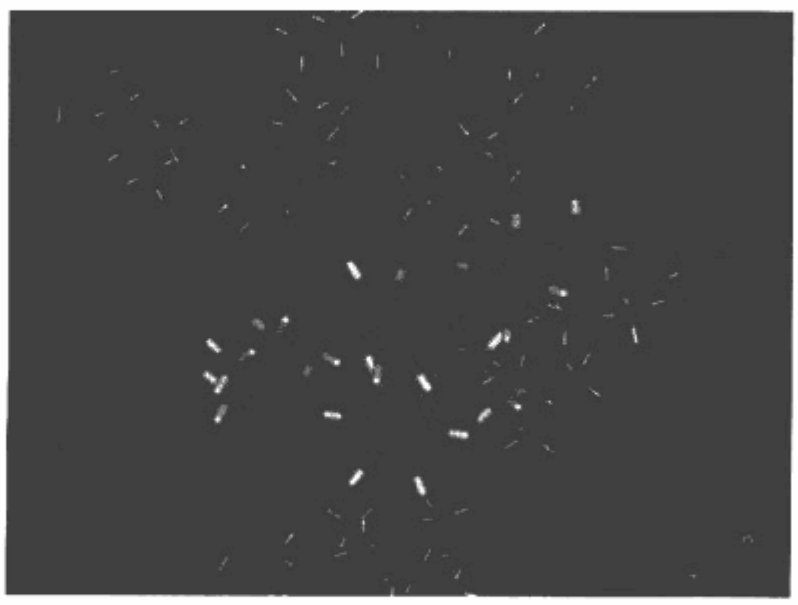


图 5-20　对接结果示意图

然后点击表格中 ↑ 键和 ↓ 键浏览对接的配体构象，随着鼠标点击表格中的不同配体对接姿态，视图窗口中的配体对接姿态会随之更新。

在 Tools 中从下拉列表中选择 Receptor-Ligand Interactions，打开 Receptor-Ligand Interactions 工具面板，在 Visualize Receptor-ligand Interactions 工具面板下点击 Receptor- ligand Hydrogen Bonds，在视图窗口中，受体原子与配体对接 poses 间的氢键会通过绿线显示出来（图 5-20）。

在表格中，继续点击剩下的配体对接姿态，这样对接姿态将被添加到系统视图中去并

逐一在图形视图中显示出来，前面 Selected Residues 参数识别出的氨基酸残基侧链构象将随着配体姿态变化而变化。

在 Visualize Receptor-ligand Interactions 工具面板下再次点击 Receptor-ligand Hydrogen Bonds，在视图窗口中，受体原子与配体对接 poses 间的氢键会取消显示。

从菜单中选择 Window | Close All，关闭所有窗口。

二、分析配体对接结果

下面我们计算配体对接构象的 RMSD 值及分析其 CDOCKER 能量。

1．在 DS 窗口中打开 3ert 配体的晶体构象

按 CTRL+1 使 Discovery Studio 浏览器恢复通常状态。在文件中找到 3ert_ligand.sd，右键点击，选择 Open With | 3D Window，3ert 配体的晶体构象将在一个新的三维窗口中打开（图 5-21）。

注意：3ert 配体已经通过分子叠合的方法放置在 1err 中，在进行分子叠合的时候，只考虑结合位点周围氨基酸残基的主链重原子。

2．在 DS 三维窗口中打开所有柔性对接产生的对接构象

在文件中，打开柔性对接作业输出文件夹，将 1err_prot.sd 拖拽到配体晶体构象所在的三维窗口，这会在同一个三维窗口打开所有的对接姿态（图 5-22）。

图 5-21　3ert 配体晶体构象

图 5-22　配体晶体结构与对接结果叠合

3．将晶体构象定义为结构比较的参考构象

在系统视图中，点击第一个配体构象（3ert_ligand）选择它，这个构象即对应晶体构象。从菜单中选择 Structure | RMSD | Set Reference，将 3ert 晶体构象设置为 RMSD 计算的参考构象。

4．计算对接产生的姿态与晶体姿态的差异

在菜单中选择 Structure | RMSD | Heavy Atoms，计算对接产生的姿态与晶体姿态的差异，RMSD 值将在新的窗口中出现（图 5-23）。

```
Heavy Atom RMSD to 3ert_ligand 1

Name                          Reference                     RMSD (A)
3ert_ligand 1                 3ert_ligand 1                     0.00
3ert_ligand 2                 3ert_ligand 1                     9.71
3ert_ligand 3                 3ert_ligand 1                     4.97
3ert_ligand 4                 3ert_ligand 1                     9.66
3ert_ligand 5                 3ert_ligand 1                     2.22
3ert_ligand 6                 3ert_ligand 1                     9.79
```

图 5-23　配体对接构象较晶体构象的 RMSD 值

第四节　利用 AutoDock 程序进行分子对接的操作及研究实例

除了前面介绍的商业对接程序之外，还有很多免费的分子对接程序。在本节里，主要讲一下目前比较流行的分子对接软件 AutoDock。AutoDock 是由 Scripps Research Institute（http://autodock.scripps.edu/）开发的一款用于预测生物大分子与其配体之间相互作用的分子对接软件，采用拉马克遗传算法。AutoDock 主要用于小分子与其受体的对接，但也可用于小肽和其受体的对接。 目前 AutoDock 的最新版本是 4.2，在 AutoDock4 之前的版本中，受体只能是蛋白，但在 AutoDock4 中，受体既可以蛋白也可以是核酸。在 AutoDock 3 中，只有几种常见的原子类型（C、H、O、N、S 等），但是在 AutoDock4 中，卤族元素、常见的金属元素等也被加了进来，大大扩展了其应用范围。以下将分别就 AutoDock4 的安装、运行及分析进行较为详细的说明。

一、AutoDock 的安装

AutoDock 的安装比较简单，主要分为两个步骤，即 AutoDock 程序包的安装和 AutoDock Tool 的安装。由于大多数计算化学研究员的工作平台都是 Linux，因此只就 AutoDock 在 Linux 平台下的安装进行说明，如果读者想在其他平台下运行 AutoDock，请参阅 AutoDock 网站上的安装指导。

1．在 Linux 下安装 AutoDock 程序包

首先登陆 AutoDock 及其他工具下载页面（http://autodock.scripps.edu/downloads），然后从该网页上部的链接进入到 AutoDock 的下载页面。在这里需要注意的一点是，下载 AutoDock 之前需要先进行信息注册，注册完后才能下载。AutoDock 支持多种操作系统，读者可以根据自己需要进行下载。另外需要提到的一点是，在 AutoDock 及其他工具的下载页面中，除了 AutoDock 程序本身，还有一个 AutoDock Tools（ADT）。顾名思义，这个 ADT 相当于一个 AutoDock 的接口软件。利用该软件，能够在图形用户界面的环境中对欲分析的体系进行各种操作，包括从受体文件、配体文件的准备到 AutoDock 的运行，再到最后的结果分析。

对于之前使用过 AutoDock 的读者来说，可能已经很熟悉 AutoDock 的工作流程了。但是对于从未接触过 AutoDock 的读者来说，在此有必要对刚刚下载的 AutoDock 文件包做一简要说明。读者下载的 AutoDock 文件包一般由两个文件组成，一个是 AutoDock，另一个是 AutoGrid。AutoGrid 负责进行分子对接之前的格点能计算（具体理论请参考 AutoDock User Guide 3），而 AutoDock 负责分子对接和结合自由能计算等任务。当获得了 AutoDock 文件包后，读者可以按照以下步骤来安装 AutoDock 文件包：

- 在 linux 系统下新建一个文件夹用来存放 AutoDock 文件包（为方便起见，作者假

设读者把该文件夹命名为 AutoDock4）

mkdir AutoDock4

• 将下载 AutoDock 文件包移动至刚刚建立的文件夹内，并且进入该文件夹

mv autodocksuite-4.2.1-i86Linux2.tar.gz / AutoDock4

cd AutoDock4

• 提取文件包中的文件，得到 AutoDock 和 AutoGrid

tar xvzf autodocksuite-4.2.1-i86Linux2.tar.gz

• 为上一步提取的 AutoDock 和 AutoGrid 添加可执行权限

chmod u+x autogrid4 autodock4

• 做环境变量影射

vi .bashrc（如果读者用的不是 bsah shell，请咨询电脑管理员）

在打开的 vi 编辑器中添加类似的环境变量

export PATH=$PATH:/home/USERNAME/AUTODOCK4_FOLDER

其中 USERNAME 和 AUTODOCK4_FOLDER 分别为用户自己的主目录名和 Autodock4 所在的文件夹。

2．在 Linux 下安装 AutoDock Tools

AutoDock Tools （ADT） 是 Scripps 研究所开发的一个集 AutoGrid、AutoDock 的输入文件的准备和结果分析等功能于一身的软件。利用该软件，AutoDock 用户可以轻松地在图形用户界面下进行各种操作。当然，对于已经很熟悉 AutoDock3 的用户，也可以使用 python 脚本进行操作。关于该软件的使用，将在本章后面的部分进行介绍。

ADT 的下载不需要注册，可直接从 http://mgltools.scripps.edu/downloads 下载，在下载时请尽量使用 FlashGet 等可以断点续传的下载工具，因为 ADT 的下载对于教育网的用户来说依然是缓慢的（笔者就曾在论坛上见过有人求助此问题）。目前 ADT 的最新版本是 1.5.4， 在下载页中，读者可以根据自己的需要下载不同版本。ADT 在 Windows 平台下的安装很简单，在此只叙述 ADT 在 Linux 平台下的安装方法。关于 ADT 在 Windows 和 Mac 平台下的安装，读者参考下在页面上相应的安装指导。以下叙述 ADT 用安装文件在 Linux 平台下安装 ADT：

• 将下载的 ADT 安装文件（rpm 格式）利用 FTP 工具（如 CuteFTP）传输到 Linux 平台下

• 新建文件夹 ADT，并将 ADT 安装文件移动到该文件夹内

mkdir ADT

mv ADT_install_file /ADT

• 给 ADT 安装文件添加执行权限

chmod u+x ADT_install_file

• 运行 ADT 安装文件

注意：如果读者是在本地 Linux 环境的 PC 机上安装 ADT 的话，可以直接运行 ADT 安装文件；如果读者是通过远程登陆在 Linux 的工作站或服务器安装 ADT 的话，需要注意不要使用 Penguin 来连接。因为运行 ADT 安装文件需要用到图形界面，而 Penguin 所建立的连接是不支持图形界面的。读者可以采用 Xmanager 等支持图形界面的远程登录软件来建立连接。

- 执行 rpm 安装包并根据提示，将 ADT 安装到目标文件夹

二、AutoDock 的基本应用

上一节叙述了 AutoDock4 在 Linux 平台下的安装过程。以下将分三个部分向读者比较系统地讲述 AutoDock 的基本应用。在完成本节的学习后，读者将能够掌握 AutoDock 的最基本应用。

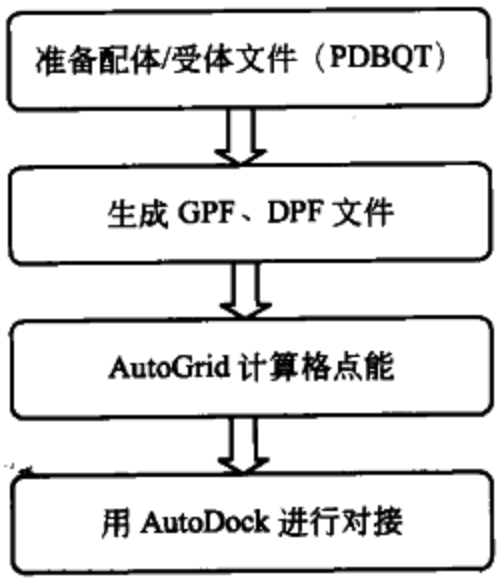


图 5-24　AutoDock 的操作流程

1．AutoDock 的操作流程

AutoDock 的操作流程如图 5-24 所示，可以分为 4 个步骤，在这 4 个步骤中，第一个步骤尤为重要，因为输入文件的正确与否直接关系到下面步骤运算结果的准确性。

2．用 AutoDock Tools 操作 AutoDock 进行分子对接

ADT 是一个可视化的分子模拟软件包，它集成了分子显示、作图、AutoDock 输入文件的准备、结果分析等功能。在这一节中，将以 HIV-1 蛋白酶（protease）及其抑制剂为例向读者介绍用 ADT 生成 PDBQT 的方法，而其分析功能将放在本章的最后加以叙述。至于其他在对接中较少涉及的功能，请有兴趣的读者参考 AutoDock 网站上的资源。此外，由于 ADT 的流行仅仅是最近一两年的事情，因此该软件在设计上还存在着一些不足，最令人费解的地方就是在生成 PDBQT 时多种多样的组合。设计者最初的目的可能是为了满足不同用户的多种多样的需求。毕竟在一次模拟过程中得到一个输出文件很可能就是其他模拟过程的输入文件，因此 ADT 有必要考虑到多种情况以便尽可能地满足用户的需要。从这一点上来说，ADT 可谓是一个非常灵活的软件，这一点是其他商业软件所无法比拟的。由于篇幅的原因，无法对每一种生成 PDBQT 的办法一一进行说明，还请读者见谅。

（1）准备配体文件和受体文件　用 AutoDock 进行分子对接的第一步是为 AutoGrid 和 AutoDock 准备其必须的配体和受体文件。在 AutoDock4 中，其配体、受体的输入文件格式为 PDBQT。PDB 文件大家应该已经很了解了，实际上其主体就是原子坐标。而 PDBQT 文件不仅包含了原子坐标，而且还包含了原子类型、部分电荷（partial charges）甚至以及柔性键等信息。

对于一个典型的分子对接运算（指在对接过程中考虑配体的部分或全部柔性，而受体保持刚性），一般需要准备两个 PDBQT 文件，一个包含配体的信息，另一个包含受体的信息。生成一个 PDBQT 文件的过程比较复杂，一般可分为以下 5 个步骤：

- 为输入的配体/受体补全 H 原子（all-atom model）
- 为所有原子添加部分电荷
- 删除非极性 H 原子，并把其部分电荷添加到其母原子（mother atom）上
- 确定原子类型，包括定义氢键受体(acceptor)/供体(donor)、定义芳环碳、脂肪族碳原子等操作
- 定义根原子（root atom）并确定柔性键

在此需要说明的一点是，读者也可以在最初为配体/受体加 H 原子时直接选择添加极性 H 原子，然后再添加电荷，这样上述前 3 个步骤可以简化为 2 个步骤，但要注意，此时不能再用 Autodock Tools 为分子添加 Gasteiger 电荷，因为 Gasteiger 电荷的计算是基于全原子模型（all-atom model）的。生成 PDBQT 文件的方法如下。

① 配体文件的准备

a. 获得配体文件　一般来说，获得配体文件有两种途径，一种是从网上数据库（如ZINC）下载，另一种是利用 ChemDraw 或 Chem3D 等软件自己构建。

b. 读入配体文件　启动 ADT，菜单栏 File | Read Molecule，在弹出的 Read Molecule 对话框中选择配体文件 ligan_1TCX.pdb，单击 OK 读入配体文件。由于 ADT 不能识别中文字符，因此不要把任何要用 ADT 处理的文件放到含有中文的文件夹或路径中（图 5-25）。此外，ADT 支持拖拽读取，当文件路径过深时，读者可利用 ADT 的此特点来快速读取数据。

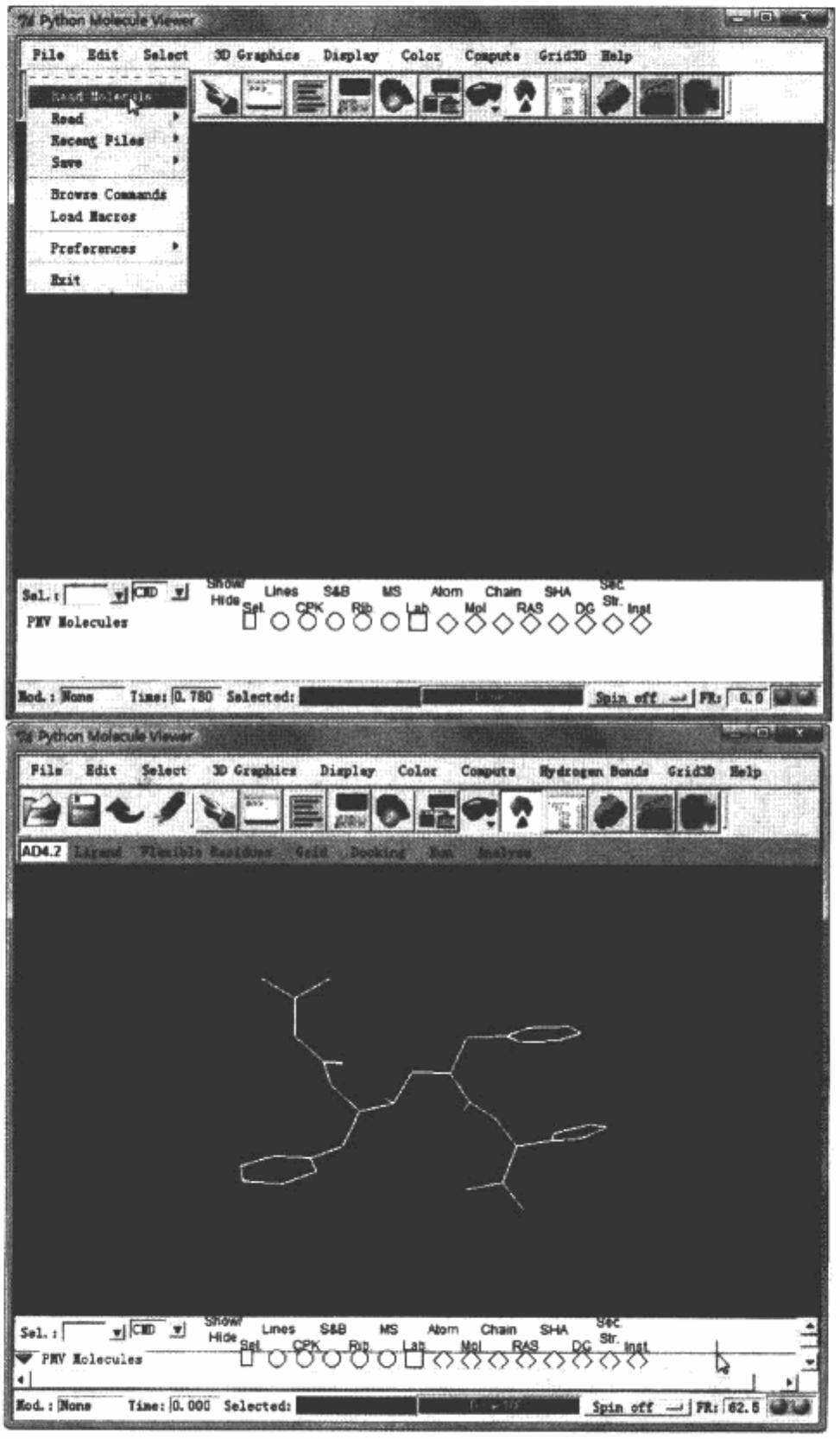


图 5-25　配体文件的读入

c. 配体加 H (all-atom model)　读入配体文件后，菜单栏 Edit | Hydrogens | Add，在弹出的 Add Hydrogens 对话框中选择 All Hydrogens、noBondOrders、yes。可选择性保存加过

H 的配体文件：File | Save | PDB（图 5-26）。

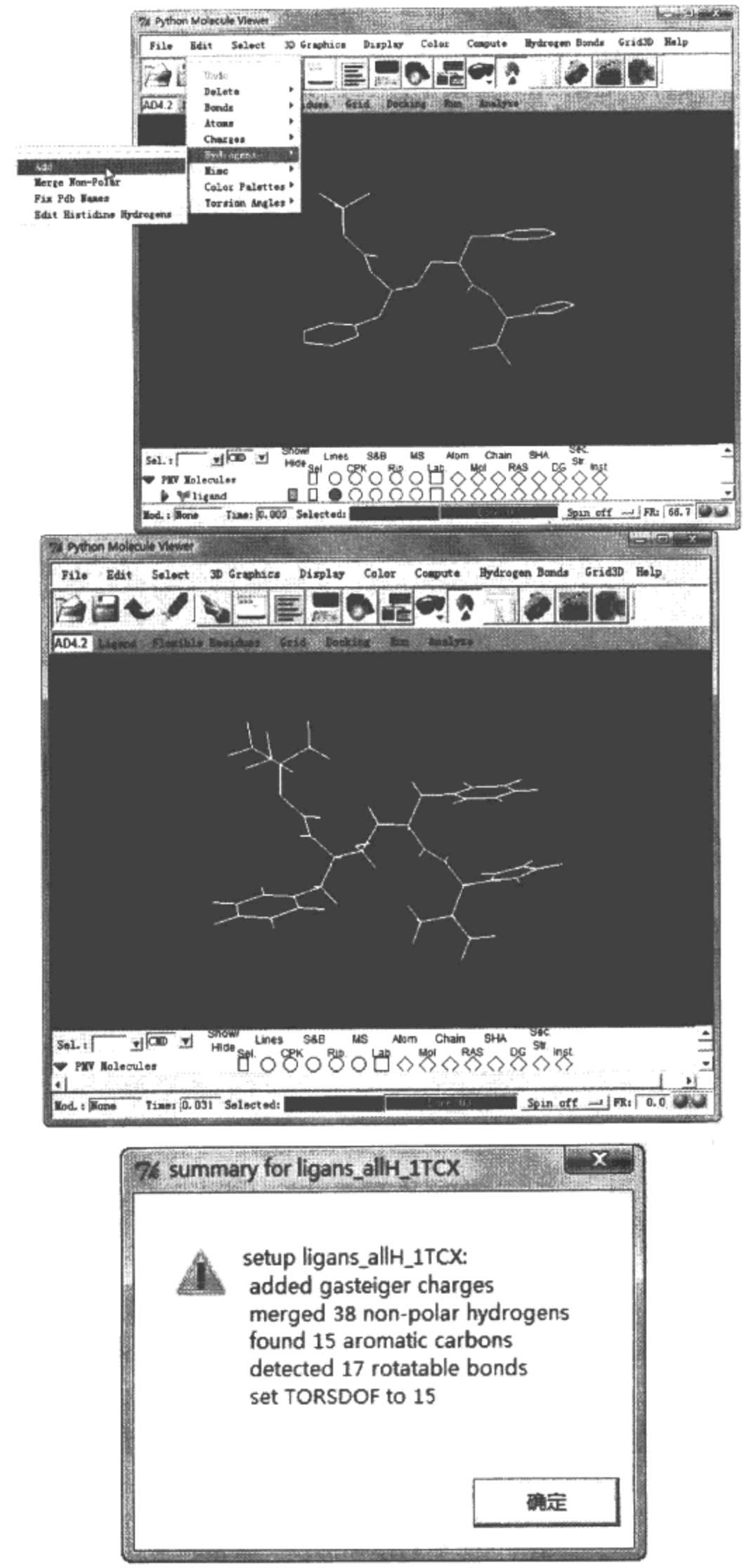
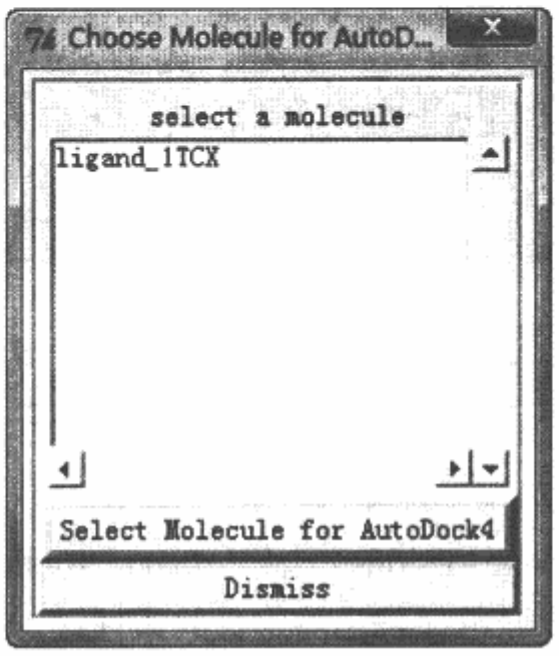


图 5-26　配体加氢

d．生成配体的 PDBQT 文件　单击工具栏上的 AutoDock Tools 标志 调出 AutoDock 菜单(图 5-27)，在 AutoDock 菜单上选择 Ligand | Input | Choose。在弹出的 Choose Molecule for AutoDock 面板上选择上一步中加过 H 的 HIV-1 整合酶抑制剂为配体分子。单击 Select Molecule for AutoDock 后大约 1s（取决于电脑速度）会弹出一个 Summary for ligand_filename （配体文件名）的对话框（图 5-28)。从这个对话框中显示的信息可以看出，实际上这一步骤是一个集成了 4 步运算的复合运算，即：

AutoDock Tools 菜单



Choose Molecule for AutoDock4 窗体

图 5-27

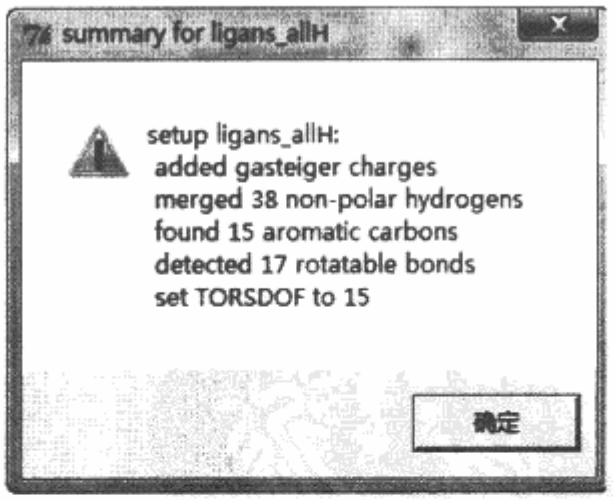


图 5-28

• 首先，ADT 检测输入的分子是否存在电荷信息，如果有则保留，如果没有则计算 Gasteiger 电荷并赋予之。这里需要注意的一点是，计算 Gasteiger 电荷需要用到全原子模型，因此要想得到正确 Gasteiger 电荷，必须保证输入的配体分子上所有的 H 原子都已经被加上。

• 其次，ADT 删除所有非极性 H 原子，并将其电荷分配到其相应的母原子上（此步骤可以通过设置 File | Preferences | Modify Default | AutoDockTools | Automerge NPHS = 0 来屏蔽掉)。

• 再次，ADT 为每一个原子分配一个 AutoDock 中定义的原子类型。A 代表芳香 C 原子，NA、OA、SA 分别代表可作为氢键受体的 N、O 和 S 原子，HD 代表作为氢键供体的 H 原子。关于具体的原子类型定义及原子类型分配规则，有兴趣的读者可以参考 AutoDock 网站上的相关资源（如 AutoDock 手册等)。除了可以让 ADT 自动分配原子类型以外，读者也可以利用 Ligand | Aromatic Carbons 来自定义芳环 C 原子。

• 最后，ADT 检测输入的配体分子的柔性并设置二面角自由度（TORSDOF，实际上就是配体上柔性键的数目)。

e. 设置根原子(root) 用 Ligand | Torsion Tree | Detect Root 让程序自动确定 root。如图 5-29 所示，root 在 3D 界面上用绿色圆球表示。

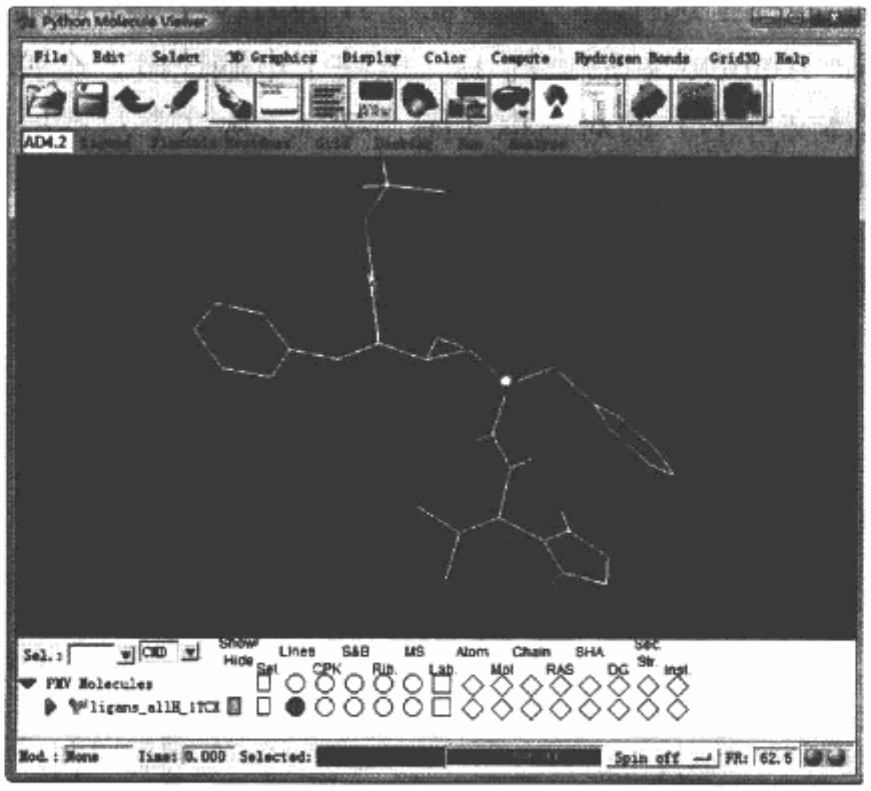


图 5-29 选择 Root

f. 手动选择柔性键（choose torsions） 用 Ligand | Torsion Tree | Choose Torsions 手动选择柔性键。如图 5-30 所示，绿色代表当前定义的柔性键（rotatable），紫色代表当前未激活的柔性键（non-rotatable），而红色代表非柔性键（un-rotatable）。在弹出的 Torsion Count 对话框中有几个按钮，读者可以通过它们快捷地设置柔性键，此外，读者还可以利用鼠标直接在 3D 窗口上选择或屏蔽柔性键。

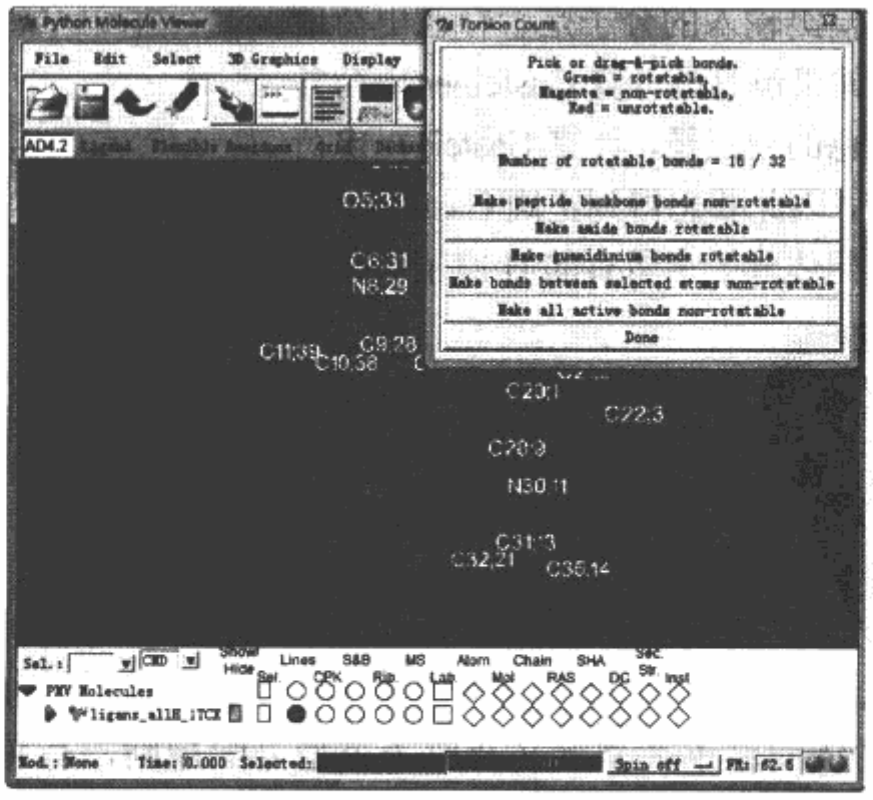


图 5-30 选择柔性键

g. 保存生成的配体 PDBQT 文件 单击 Ligand | Output | Save as PDBQT 以保存生成的配体 PDBQT 文件。

② 刚性受体文件的准备

a. 获得受体文件　到 RCSB Protein Data Bank 下载研究的受体结构。

b. 检查并完善受体结构　首先利用 VMD（Visual Molecular Dynamics）或 ADT 等可视化分子模拟工具检查受体结构是否完整，然后删除结晶水、氢原子及其他在结晶中加入的离子等物质(具体方法请参考 VMD 手册或者 AutoDock 网站上的 Using AtutoDock4 with ADT)，然后将处理好的受体另存为一个 PDB 文件。

c. 读入受体文件　用 File | Read Molecule 读取之前处理好的受体文件。

d. 受体分子加 H　读入受体分子后，用 Edit | Hydrogens | Add，在弹出的 Add Hydrogens 对话框中选择 All Hydrogens、noBondOrders、yes。可以选择性保存加过 H 之后的刚性受体文件，File | Save | PDB。

e. 生成并保存刚性受体 PDBQT 文件　单击工具栏上的 AutoDock Tools 标志调出 AutoDock 菜单，在 AutoDock 菜单上选择 Grid | MacroMocule | Choose，在弹出 Choose Macromolecule 对话框（图 5-31）中选择欲处理的受体文件(注意，受体文件也加全 H)并确定。经过大概 1s 的运算，会弹出一个警告对话框，如图 5-32 所示，显示程序在输入的受体文件中发现了 1284 个非极性 H 原子，并已将它们融合到其母原子上。点击 OK，在弹出的 Modified AutoDock4 Macromolecule File 中选择欲存储的文件名和路径以保存刚性受体的 PDBQT 文件。实际上，与生成配体的 PDBQT 文件时类似，这一步也包含了多步运算：

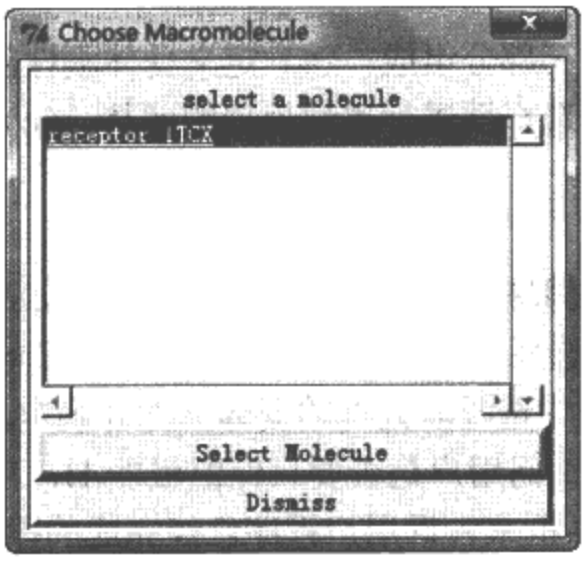


图 5-31　选择受体

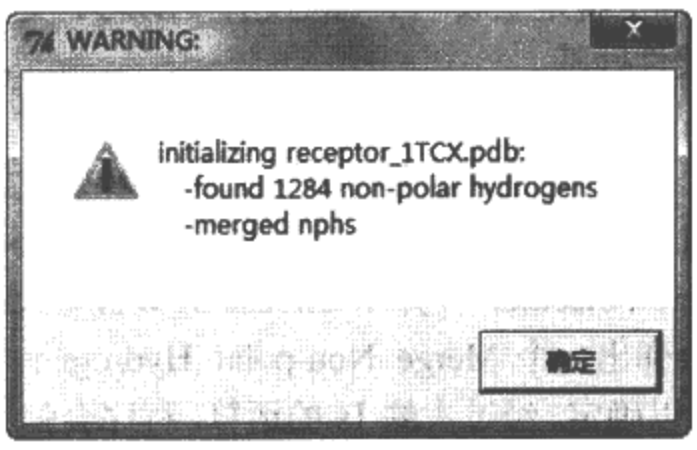


图 5-32　结果对话框

• 首先，ADT 检测输入的分子是否存在电荷信息，如果有则可以选择保留，如果没有则计算 Gasteiger 电荷并赋予之。这里需要注意的一点是，计算 Gasteiger 电荷需要用到全原子模型，因此要想得到正确 Gasteiger 电荷，必须保证输入的配体分子上所有的 H 原子都已经被加上。

• 其次，ADT 删除所有非极性 H 原子，并将其电荷分配到其相应的母原子上（此步骤可以通过设置 File | Preferences | Modify Default | AutoDockTools | Automerge NPHS = 0 来屏蔽掉）。

• 最后，ADT 为受体中的每一个原子分配原子类型并生成结果报告。

③ 半柔性受体文件的准备　在 AutoDock3 中，受体分子被视为完全刚性，即使是受体活性口袋区域的柔性也没有进行任何特殊的处理。然而受体的柔性在很大程度上决定着

其与底物的结合方式， 一个典型的例子是细菌质膜外结合蛋白（periplasmic binding proteins）。尽管在一级结构上差别很大，但是这一类蛋白却具有一种非常相近的三级结构，即由一个连接结构域（interconnecting domain）和两个球状结构域（globular domain）构成，两个球状结构域位于两侧，中间由连接结构域相连。此类蛋白一般游离于革兰阴性菌的内膜与外膜之间，通过其特殊的结构与功能来捕获外界进入的营养物质，从而供机体吸收。大量的实验和模拟结果已经证明，在捕获其底物的过程中，质膜外结合蛋白连接结构域向外弯曲，两个球状结构域相互靠近，从而最终将其底物牢牢地困在其结合口袋中。由于质膜外结合蛋白的球状结构域的大规模运动很像捕蝇草捕捉小昆虫时的过程，因此形象地将质膜外结合蛋白的这种工作机理称为“捕蝇草工作机理”。

通过上面的背景知识可以看到，在很多情况下，仅仅考虑配体的柔性是远远不够的。希望能有一种方法可以合理地、快捷地处理受体的柔性问题，但是到目前为止，受体的柔性问题依然没有得到很好的解决。目前主要有两种考虑受体柔性的方法。一种是 MD 模拟和分子对接联用的方法，另一种是在分子对接的过程中考虑受体部分残基侧链柔性的方法。在这里，主要讲解在 AutoDock4 中考虑受体侧链柔性的方法。用 ADT 生成柔性受体 PDBQT 文件的步骤如下。

a. 读入受体文件 用 File | Read Molecule 读取之前处理好的受体文件。

b. 受体分子加 H 读入受体分子后，用 Edit | Hydrogens | Add，在弹出的 Add Hydrogens 对话框中选择 All Hydrogens、noBondOrders 或 yes。可以选择性保存加过 H 之后的刚性受体文件，File | Save | PDB。

c. 生成半柔性受体 PDBQT 文件 此步操作由三部分组成：（a）定义受体刚性部分（实际上就是生成一个刚性受体 PDBQT 文件）；（b）选择欲进行柔性处理的残基；（c）定义柔性残基。

- 定义受体刚性部分

这一步与前面生成刚性受体 PDBQT 文件的过程很相似。单击工具栏上的 AutoDock Tools 标志调出 AutoDock 菜单，在 AutoDock 菜单上选择 Flexible Residues | Input | Choose Macromolecule，在弹出 Choose Macromolecule 对话框中选择欲处理的受体文件并确定。此时会弹出一个 Merge Non-polar Hydrogens 对话框来询问用户是否要只保留极性 H，单击 Yes 以确定。经过大约 1s 的运算，程序会弹出一个对话框来告诉用户程序已经完成的操作。

- 选择柔性残基

ADT 为用户提供了强大的“选择工具”，在此向读者介绍两种较常用的选择操作方法，读者可以根据自己的操作习惯选取其中的一种来进行选择操作。

鼠标法：按住 shift+鼠标左键在 3D 窗口内直接选择欲进行柔性处理的残基中的某个原子，然后打开菜单栏中 Select | Set Selection Level 菜单，设置 Selection level 为 Residue（图 5-33）。此时读者会看到刚才选择的点（黄色叉子）已经扩展到了整个残基上。如果读者想连续选择多个柔性残基，只需要在选择过程中按住 shift 键不放，然后用鼠标左键点击欲选择的原子即可。

菜单法：打开 Select 菜单，读者可以看到它包含了很多的子菜单，选择其中的 Select From String，读者可以在弹出的对话框中填上各种字符串组合来进行选择。假如选择残基号为 1 的 Pro，那么就应该在对话框的 Residue 这一行的文本框中填入 PRO1（注意必须为大写），然后点击 Add。关于 Select From String 的具体语法读者可以参考 ADT 手册（图 5-34）。

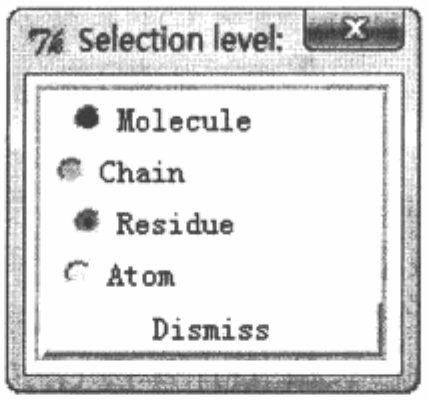


图 5-33　鼠标法选择柔性残基

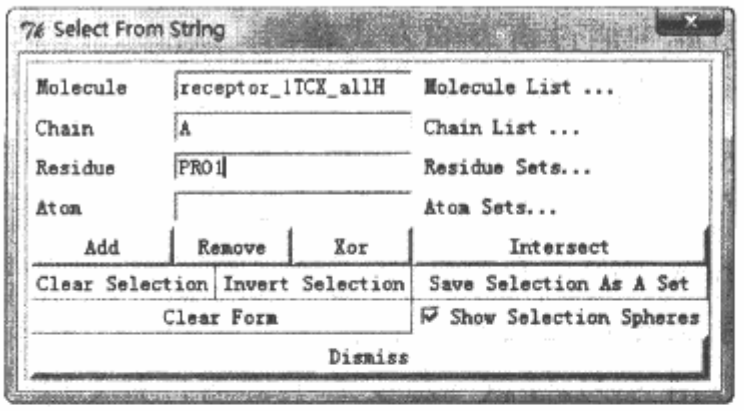


图 5-34　菜单法选择柔性残基

- 定义柔性残基

选择 Flexible Residues | Choose Torsions in Currently Selected Residues。完成如上操作后，程序会隐藏所有未被选择的残基而只留下欲处理的残基。此外，程序还会弹出一个 Torsion Count 对话框。与设置配体柔性键时类似，在这里 ADT 允许用户激活或屏蔽程序默认的柔性键。确定好柔性键后单击 Close 以退出（图 5-35）。

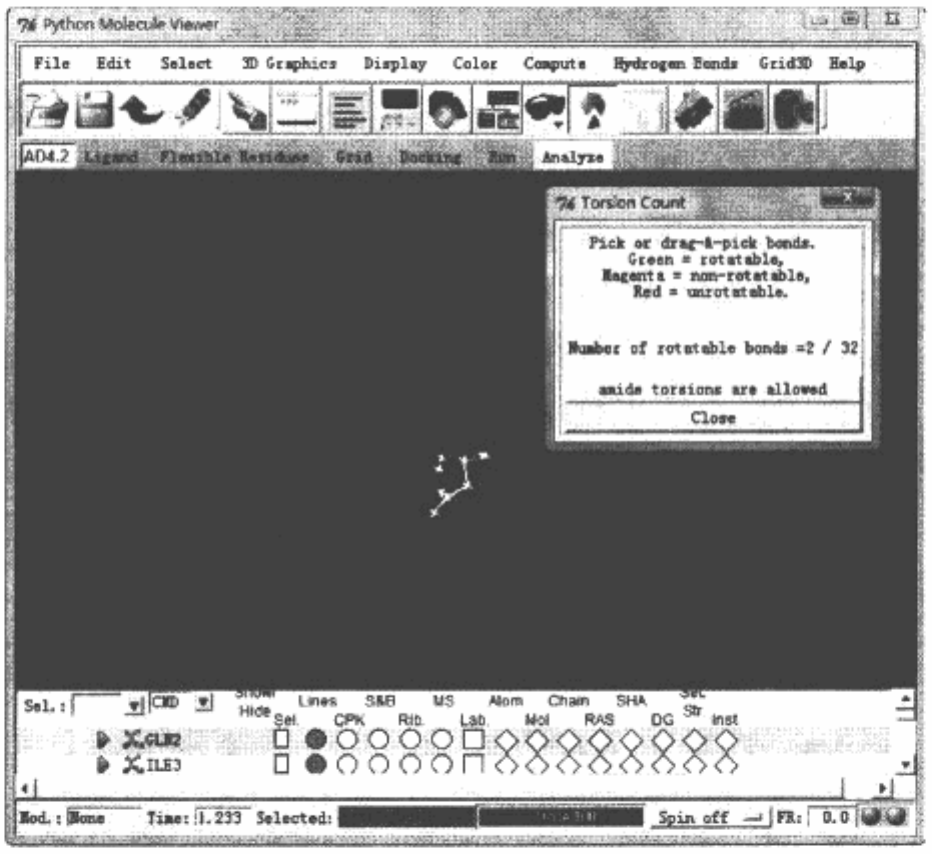


图 5-35　定义柔性残基

d. 保存生成的配体 PDBQT 文件　做柔性受体对接需要两个受体文件，一个是刚性受体 PDBQT 文件，一个是柔性残基 PDBQT 文件。因此与刚性受体文件准备过程不同是，在这里需要分别存储上述两个受体 PDBQT 文件：

单击 Flexible Residues | Output | Save Flexible PDBQT 以保存柔性残基 PDBQT 文件。

单击 Flexible Residues | Output | Save Rigid PDBQT 以保存刚性受体 PDBQT 文件。

（2）为 AutoGrid 的生成输入文件 Grid Parameter File　在进行 AutoGrid 计算之前，必须生成其输入文件，AutoGrid 唯一的输入文件是 Grid Parameter File，即所谓的 GPF。GPF 中包含了所有的 AutoGrid 所需要的信息，比如要计算格点能的受体、配体以及格点中心、范围等。AutoGrid 会为配体当中每一种原子类型计算出一个 map。这些 map 相当于能量参数库，用 AutoDock 进行分子对接运算时用到的任何配体-受体之间的相互作用能都可以从

这些能量库中通过查表的方式获得，从而大大地节省了计算时间。除了上述 map 外，ADT 还会另外计算两个 map 来共 AutoDock 使用，一种是静电 map，另一种是去容积化效应 map。通过以上叙述可见，map 的数量是可以预先确定的，比如配体（或受体上定义柔性残基）有 N 种原子类型（注意，这里说的原子类型是指 AutoDock 中规定的原子类型，而不是指 pdb 中的元素类型），那么生成的 map 的数目就应该是 N+2。利用这个关系，可以粗略判断 AutoGrid 是否工作正常。此外，GPF 实际上是一个文本文档，在生成 GPF 后，读者可以用 ADT 任意一款文本编译器打开 GPF 对其进行修改。用 ADT 生成 GPF 的方法如下。

① 生成刚性对接的 GPF 文件　首先，读入之前生成的配体、受体的 PDBQT 文件，然后单击工具栏上的 AutoDock Tools 标志调出 AutoDock 菜单，在 AutoDock 菜单上选择 Grid | Macromolecule | Choose/Open 来选择或打开受体的 PDBQT 文件。之后会弹出一个对话框说此受体文件已经包含了电荷信息，问是否保留，这里一定要选择保留，否则程序会用当前的受体结构重新计算电荷。由于现在的受体已经是联合原子模型（united-atom model）的了，所以不能用于 Gasteiger 电荷的计算（此外，程序还可弹出另外几个对话框，一般都可以忽略掉，因为 GPF 实际上是一个很简单的文件，它只告诉 AutoGrid 如何去计算 map，并不包含配体或受体的实际信息）。

其次，用 Grid | Set Map Types | Choose/Open Ligand 在弹出的对话框中选择配体文件。

再次，用 Grid | Grid Box 打开 Grid Option 工具（图 5-36）对每个方向上格点的数目、盒子的中心以及格点间距进行调解。盒子的大小由各方向上格点的数目和格点间距共同决定。一般来说，盒子的大小要至少能使得欲对接的配体分子即使在其最伸展的状态下也能在盒子内自由转动。盒子中心一般选择在活性位点中心。格点间距一般不用修改，这个值越大则格点越密集，计算的格点能也越准确，但是计算时间也会相应增长。在修改各个参数的同时，在 3D 窗口中读者可以实时观察到各个参数对盒子的大小及位置的影响。修改完相应的参数之后，通过 Grid Option 工具的菜单栏 File | Close Saving Current 来关闭此工具并保存参数。

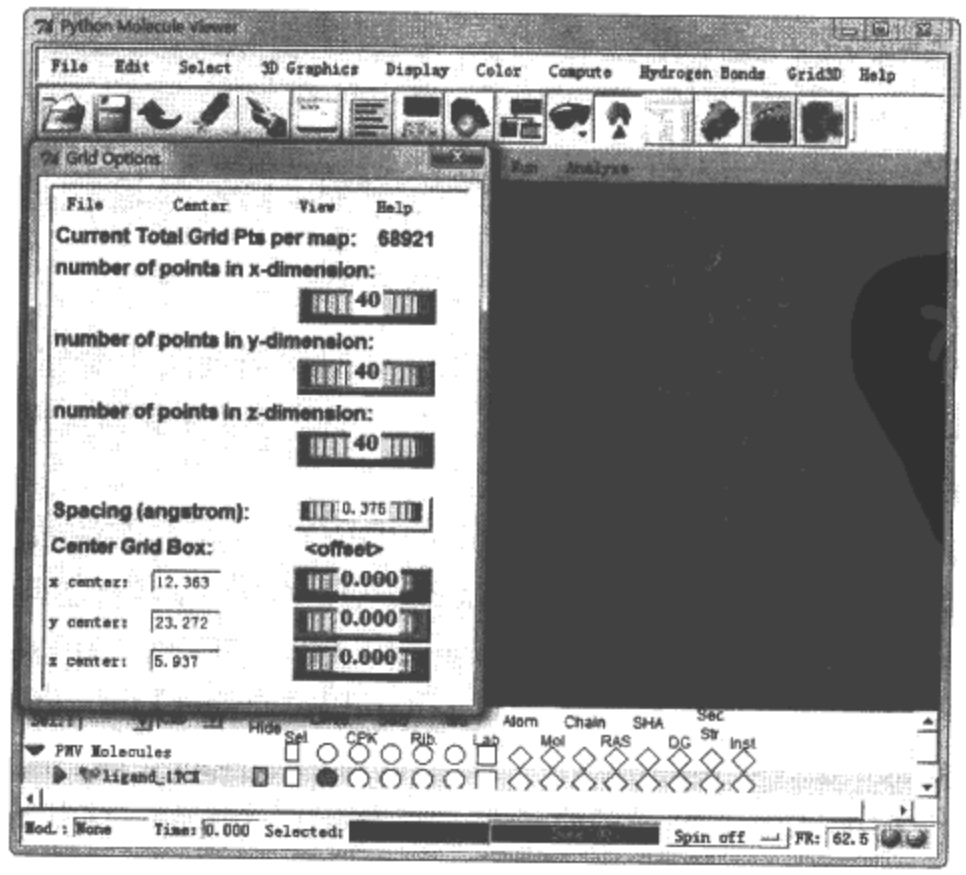


图 5-36　生成 GPF 文件

保存 GPF。用 Grid | Output | Save GPF 来保存 GPF。

② 生成半柔性对接的GPF文件　为半柔性对接生成GPF文件的方法与为全刚性对接生成 GPF 文件的方法一样，但是需要注意的是，在用 Grid | Macromolecule | Choose/Open 来定义刚性受体的时候要选择受体刚性部分的 PDBQT 文件，而不要用整个受体的 PDBQT 文件作为输入。读者只要时刻记住 AutoGrid 格点能计算是针对刚性受体和配体而进行的，AutoGrid 不能用于柔性残基和配体的格点能计算。

（3）为 AutoDock 的生成输入文件 Dock Parameter File (DPF)　DPF 中包含了 AutoDock 所需要的一切信息，如在对接中需要用到的 map、对接中要用到的配体、受体所在的文件名及位置、各种控制 AutoDock 运算时间与性能的参数等。

在 AutoDock4 中，一共有 4 种不同的对接算法，分别是 SA、GA、LS 和 GALS（LGA），其中：SA——最初的蒙特卡罗模拟退火法（Monte Carlo simulated annealing）；GA——传统的达尔文遗传算法（Darwinian genetic algorithm）；LS——局部搜索算法（local search）；GALS（LGA）是 GA 和 LS 的组合算法，也叫做拉马克遗传算法（Larmarckian genetic algorithm）。

每一组算法都有它们自己的一套参数，一般情况下不需要做修改。与 GPF 一样，DPF 也是一个文本文件。在生成 DPF 后，读者可以用任意一款文本编译器打开 DPF 对其进行修改。

① 选择刚性受体　单击 AutoDock 图标调出 AutoDock Tools 菜单，选择 Docking | Macromolecule | Set Rigid Filename，在弹出的对话框里选择刚性受体 PDBQT 文件并确定。

注意：该操作并不会读入受体文件，而只是告诉程序刚性受体的文件名。

② 选择柔性残基文件（可选）　在 AutoDock Tools 菜单中，选择 Docking | Macromolecule | Set Flexible Residues Filename，在弹出的对话框里选择刚性受体 PDBQT 文件并确定。同样地，该操作并不会读入柔性残基文件，而只是告诉程序刚性受体的文件名。此外，如果选择进行柔性受体对接的话，在最后生成的 DPF 中会出现一个 flexres 参数来告诉 AutoDock 柔性残基文件的位置，而在刚性对接中则没有该参数。

③ 读入配体文件　首先将配体的 PDBQT 文件读入（File | Read Molecule），然后在 AutoDock Tools 菜单中，选择 Docking | Ligand | Choose，在弹出的对话框里选择配体的 PDBQT 文件并确定。在这里需要注意一点，虽然 Docking | Ligand 菜单中也有 Open 子菜单，但是，通过该命令打开配体文件的 PDBQT 后程序经常会报错，并且导致无法存储 DPF 文件。所以在这里最好的解决办法是读入配体的 PDBQT 文件，然后用 Docking | Ligand | Choose 来选择配体。

④ 设置对接参数　在 AutoDock Tools 菜单中，选择 Docking | Search Parameter | Genetic Algorithm 以调出 Genetic Algorithm Parameter 编辑窗口（读者也可以根据自己需要选择其他算法）。如图 5-37 所示，读者可以自由编辑这些参数。其中，Number of GA Runs 代表着要执行的对接运算的次数，每一次完整对接运算都会得到一个结合模式；Maximum Number of evals 表示在一次对接中最大的运算次数。在实际应用中，这两个参数有时需要做些调整。比如，为了得到较合理的聚类分析，Number of GA Runs 通常会设置为 100 或更大；而 Maximum Number of evals 则要根据之前定义的柔性键的数目作相应的调整，柔性键越多，这个值越大。除了图中所示的这些参数外，实际上用户还可以用 Docking | Edit

DPF 或其他文本编辑器打开生成的 DPF 来进行更细致的编辑。

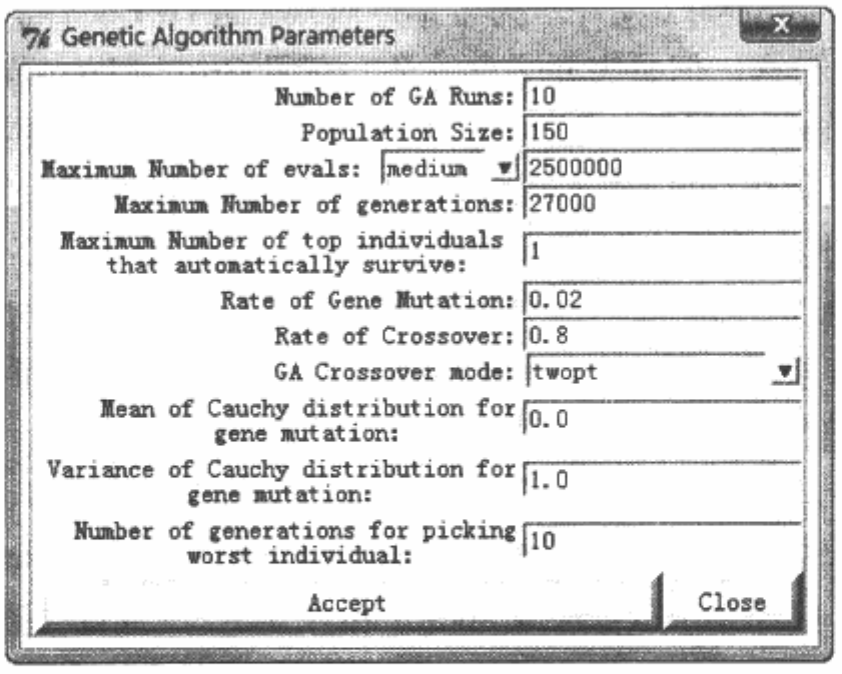


图 5-37　对接参数的设置

⑤ 保存 DPF　在 AutoDock Tools 菜单中，选择 Docking | Output | Lamarckian GA（4.2）保存生成的 DPF。

（4）运行 AutoGrid　由于 GPF 中参数不能包含路径，因此在运行 AutoGrid 之前，首先必须确保配体、受体、柔性残基的 PDBQT 文件、GPF 文件在同一个文件夹下。AutoGrid 的运行过程如下。

① 设置 AutoGrid 运行参数　在 AutoDock Tools 菜单中，选择 Run | Run AutoGrid 调出 Run AutoGrid 窗体。如图 5-38 所示，用户需要在相应的位置填入进行计算的计算机名（默认为 ADT 所在的计算机）、AutoGrid4 程序所在的路径、GPF 的文文件名及路径、输出 AutoGrid 的 log 文件（GLG 格式）的路径等参数。

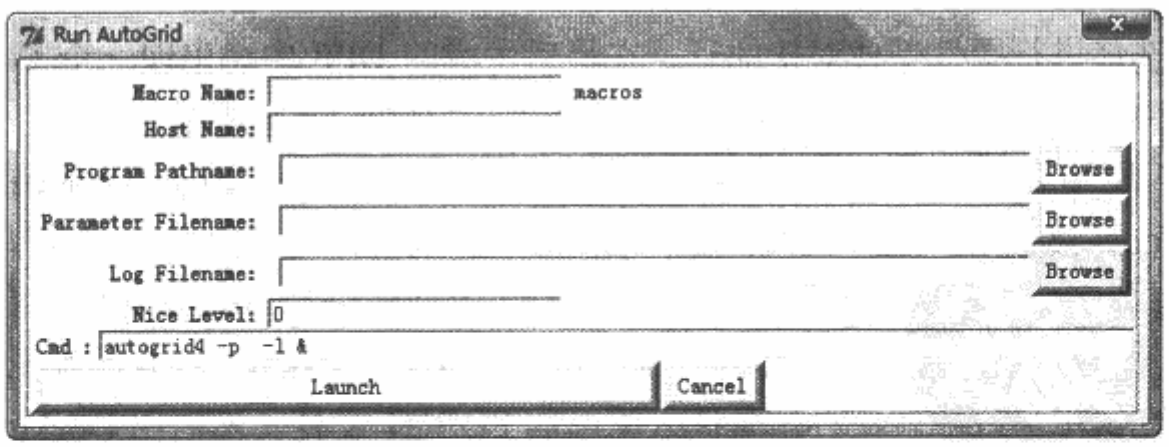


图 5-38　AutoGrid 运行参数的设置

② 启动 AutoGrid　设置好 AutoGrid 的参数后，点击 Launch 运行 AutoGrid。用户可以在 Linux 提示符下输入 top 命令并回车以查看 AutoGrid 的运行情况。

（5）运行 AutoDock　与运行 AutoGrid 一样，由于 DPF 中参数不能包含路径，因此在运行 AutoDock 之前，首先必须确保配体、受体、柔性残基的 PDBQT 文件、DPF 文件在同一个文件夹下。AutoDock 的运行过程如下。

① 设置 AutoDock 运行参数　在 AutoDock Tools 菜单中，选择 Run | Run AutoDock 调出 Run AutoDock 窗体。如图 5-39 所示，用户需要在相应的位置填入进行计算的计算机

名（默认为 ADT 所在的计算机）、AutoDock4 程序所在的路径、DPF 的文文件名及路径、输出 AutoDock 的 log 文件（DLG 格式）的路径等参数。

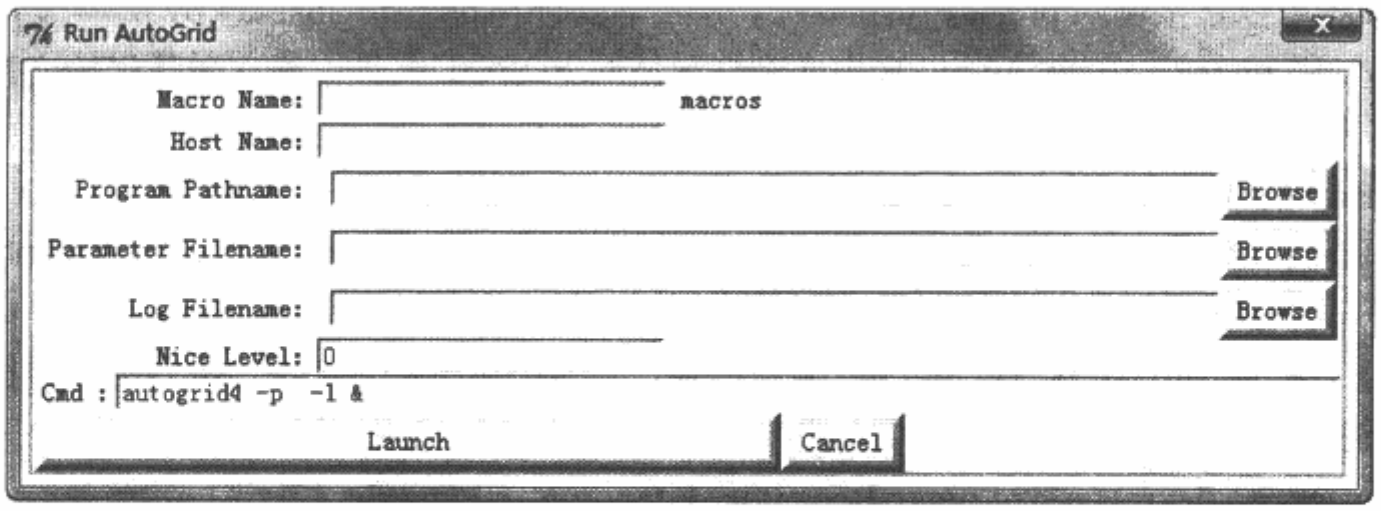


图 5-39　AutoDock 运行参数的设置

② 启动 AutoDock　设置好 AutoDock 的参数后，点击 Launch 运行 AutoDock。用户可以在 Linux 提示符下输入 top 命令并回车以查看 AutoDock 的运行情况。

3．用 Python 脚本操作 AutoDock 进行分子对接

以上着重讲解了用 ADT 操作 AutoGrid、AutoDock 进行分子对接的方法。从上面的叙述可以看出，ADT 为用户提供了可视化的操作环境，可使初次接触 AutoDock4 的用户很快学会用 AutoDock 对接的方法，然而，对于 AutoDock 的老用户来说，或许还是更加青睐用脚本语言来操作 AutoDock。虽然在 AutoDock4 的手册中并没有叙述如何用脚本来进行对接，但实际上在安装 ADT 的时候，就已经安装了运行 AutoDock 所需的所有 Python 脚本程序。因此，对于 AutoDock 的老用户来说，依然可以通过写脚本的方式来控制 AutoGrid 和 AutoDock 的运行。

（1）为 Python 脚本设置环境变量（可选）　在用 Python 脚本进行操作前，最好为其添加环境变量，否则每一次操作都要为其输入相应的路径，不仅非常繁琐，而且容易出错。为 ADT 的 Python 脚本添加环境变量的方法如下。

① 打开.bashrc 并编辑

vi .bashrc

② 在.bashrc 中添加如下环境变量

export

PATH=$PATH:/home/**USERNAME/ADTPATH**/MGLTools-1.5.4/

MGLToolsPckgs/AutoDockTools/Utilities24

其中 USERNAME 为用户主目录，ADTPATH 为 ADT 安装目录。

（2）用 Python 脚本生成配体、受体的 PDBQT 文件　相对而言，用 Python 生成 PDBQT 文件的过程较简单，一般只需要一行命令就可以完成配体或受体的 PDBQT 文件的准备。生成配体、受体及柔性残基 PDBQT 文件的 Python 脚本分别是 prepare_ligand4.py、prepare_receptor4.py 和 prepare_flexreceptor4.py。下面介绍这三个脚本的使用方法。

① 用 prepare_ligand4.py 来生成配体的 PDBQT 文件　首先将原始的配体、受体坐标文件（pdb、mol2 或 pdbq 格式）拷贝到同一个文件夹中，然后进入该文件夹，执行如下命令：

prepare_ligand4.py -l Ligand_Filename -U nphs_lps

其中 Ligand_Filename 为配体坐标文件，nphs_lps 参数说明将配体分子上的非极性 H 原子

和孤对电子（lone pair）“融合（merge）”其母原子上。实际上，prepare_ligand4.py 脚本包含了很多可选择的选项，用户可以根据自己的需要将不同的参数控制项添加到上述命令行中。读者可以直接输入 prepare_ligand4.py 并回车以查看 prepare_ligand4.py 的参数列表。

② 用 prepare_receptor4.py 来生成刚性受体的 PDBQT 文件　进入受体坐标文件所在的文件夹，执行如下命令：

prepare_receptor4.py –r Receptor_Filename -U nphs_lps

其中 Receptor_Filename 为受体坐标文件。同样读者也可以直接输入 prepare_receptor4.py 并回车以查看 prepare_ receptor4.py 的参数列表。

③ 用 prepare_ flexreceptor4.py 来生成柔性残基的 PDBQT 文件　进入受体坐标文件所在的文件夹，执行如下命令：

prepare_ flexreceptor4.py –r Receptor_Filename –s Flexible_List -U nphs_lps

其中 Receptor_Filename 为受体坐标文件，Flexible_List 为柔性残基列表，其语法为：

分子名：链号：残基全称，分子名：链号：残基全称，……

例如：hsg1: A: ARG1, hsg1: A: LYS5, ……(注意，残基名要大写)

同样读者也可以直接输入 prepare_ flexreceptor4.py 并回车以查看 prepare_ flexreceptor4.py 的参数列表。

（3）用 Python 脚本生成 GPF 和 DPF　生成 GPF 和 DPF 的 Python 脚本分别是 prepare_gpf4.py 和 prepare_dpf4.py。下面分别介绍这两个脚本的应用。

① 用 prepare_ gpf4.py 来生成 GPF 文件　首先将确保配体、受体的 PDBQT 文件在同一个文件夹中，然后进入该文件夹，执行如下命令：

prepare_ gpf4.py -l Ligand_PDBQT –r Receptor_PDBQT

其中 Ligand_PDBQT 和 Receptor_PDBQT 分别为配体与受体的 PDBQT 文件。

注意：如果要为半柔性对接生成 GPF 文件的话，则 Receptor_PDBQT 一定要选择受体刚性部分的 PDBQT 文件。

② 用 prepare_ dpf4.py 来生成 DPF 文件　首先确保配体、受体的 PDBQT 文件在同一个文件夹中，然后进入该文件夹，执行如下命令：

prepare_ dpf4.py -l Ligand_PDBQT –r Receptor_PDBQT

其中 Ligand_PDBQT 和 Receptor_PDBQT 分别为配体与受体的 PDBQT 文件。

（4）用命令行运行 AutoGrid 和 AutoDock　读者可以用如下的命令分别运行 AutoGrid4 和 AutoDock4：

autogrid4 -p GPF -l GLG

autodock4 -p DPF-l -l DLG

其中 GPF、DPF、GLG 和 DLG 分别为 AutoGrid 参数文件、AutoDock 参数文件、AutoGrid 输出文件和 AutoDock 输出文件。这里需要注意的一点是，一定要等到 AutoGrid 计算完成后才能运行 AutoDock，而且还要确保配体、受体即柔性残基的 PDBQT 文件和 AutoGrid 生成的全部 map 文件在同一文件夹下。

由于 AutoDock 运算时间一般比较长，因此可以在后台提交作业，命令如下：

nohup autodock4 -p DPF-l -l DLG

4．AutoDock 的结果分析

（1）评价对接结果的原则　一般来说，评价对接结果的优劣一般有三个原则，即结果

打分（score）、聚类（clustering）及化学合理性（chemical reasonableness）。

① 打分　对对接结果的打分是对对接优劣的最简单、最直接的评价。在 AutoDock 中，对接结结果给出的结合自由能越低，则说明对接得到的复合物的结合模式最牢固。因此，在分子对接中，一般来说，一个打分较高（能量较低）的构象一般都比一个打分较低（能量较高）的构象要更有说服力。然而，打分评价标准只有在两种构象的打分相差较大的情况下才有意义。拿 AuotoDock 来说，其结合自由能的计算误差在2kcal/mol 左右，因此不能说一个结合自由能为–5kcal/mol 的结合模式要优于一个–4kcal/mol 的结合模式。此外，由于结合自由能最低的结合模式未必就是天然状态下的结合模式，所以，即使得到一个能量明显优于其他结合模式的结果，也不能确定它就是正确的结合模式。由以上叙述可见，仅仅依靠对接打分来评价对接结果显然是不够的。

② 聚类　聚类是指在一定的阈值（cutoff）下，将对接结果按照结构和能量上的差异进行归类的一种统计学方法。在 AutoDock 聚类分析中，程序会首先按结合自由能从低到高将对接结果排序，然后再按照给定的均方根偏差（RMSD）阈值来对排序好的对接结果进行分类。可以想象，一种占优势的结合模式在分子对接计算中被发现的概率要远大于那些非优势结合模式。因此，有理由相信，所要寻找的结合模式应该位于那些聚类元素较多的类当中。在一次成功的对接中，其聚类的数目不应过多，否则会使得每一类中的元素数目较少；此外，位于能量较低类的元素数目应明显多于位于能量较高类的元素数目。一般来说，如果能找到一种结合模式，其能量较低、聚类又很好的话，那么该结合模式就比较可能是正确的结合模式。

③ 化学合理性　化学合理性原则就是在各种分子模拟软件的帮助下，依靠个人的化学直觉对通过上面两种标准选择出来的结合模式进行细致的观察、分析。一般需要检查的细节有：氢键的数目与形成位置、疏水性集团之间的相互作用、盐桥等。当然，如果在对接前就通过阅读文献或实验结果已经知道了部分结合模式的信息，如配体的整体走向，那么必须把这些信息当作判断正确结合模式的首要标准。

（2）用 ADT 分析 AutoDock 结果　做分子对接的目的就是要预测配体与受体的结合模式。上面已经讨论了选择合理结合模式的标准，以下简单介绍 ADT 中的分析工具，并就几个大家可能比较关心的内容展开讨论。

① 读入 DLG 文件　用 ADT 分析对接结果的第一步就是将 DLG 文件导入到 ADT 中，首先在 ADT 的工具栏中点击 AutoDock 图标调出 AutoDock Tools 菜单，然后选择 Analyze | Docking | Open 来打开欲进行分析的 DLG 文件。此时，ADT 会弹出一个类似于图 5-40 的对话框，大意是告诉用户读入的文件名及构象数目，并且告诉用户可以用 Analyze | Conformations | Play 菜单来观察每一种结合模式。点击确定，关闭该对话框。

② 获得每种对接模式的基本信息　在完成一次对接计算之后，都要检测每种结合模式的各种指标，从而对对接结果进行评价。可以通过文版编辑器打开 DLG 文件，然后自己去搜索每一种结合模式的各种信息。除了这种方法外，ADT 还提供给一种可视化的手段来检测每一种结合模式的各种信息。

选择 Analyze | Conformations | Load 来打开 Conformation Chooser 窗体。如图 5-41 所示，该窗体包含上下两个部分，下半部分列出了所有的结合模式的聚类及结合自由能信息，而上半部则列出了相应结合模式的详细信息，如打分排序（rank）、结合和自由能（binding energy）等各种能量、常温下结合常数、与参考构象的 RMSD（Ref RMS）、与本类中能量

最低构象的 RMSD（Cluster RMS）等。这个工具除了可以显示不同构象的详细信息之外，还可以在下半部列表中通过双击任意一行的方法来显示其相应的 3D 结构。为了方便分析，通常还需要把受体也显示在 3D 窗口中，这一点可以通过 Analyze | Macromolecule | Open 或 Choose 来实现。

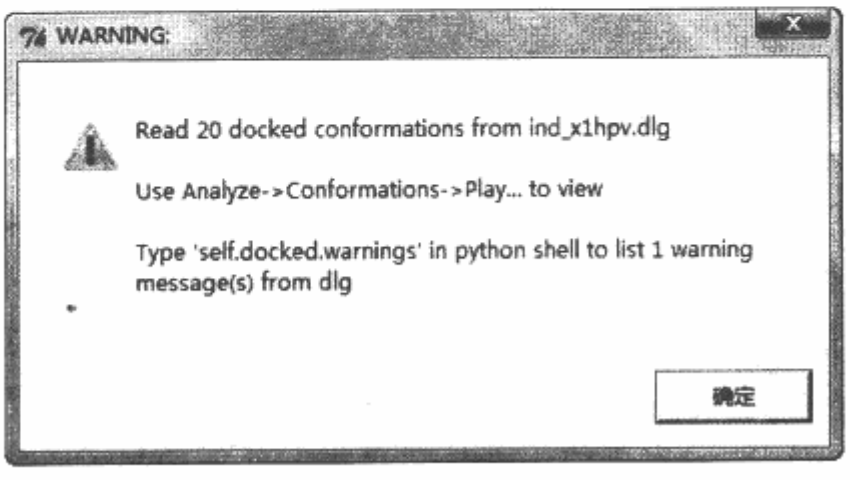


图 5-40　DLG 文件导入

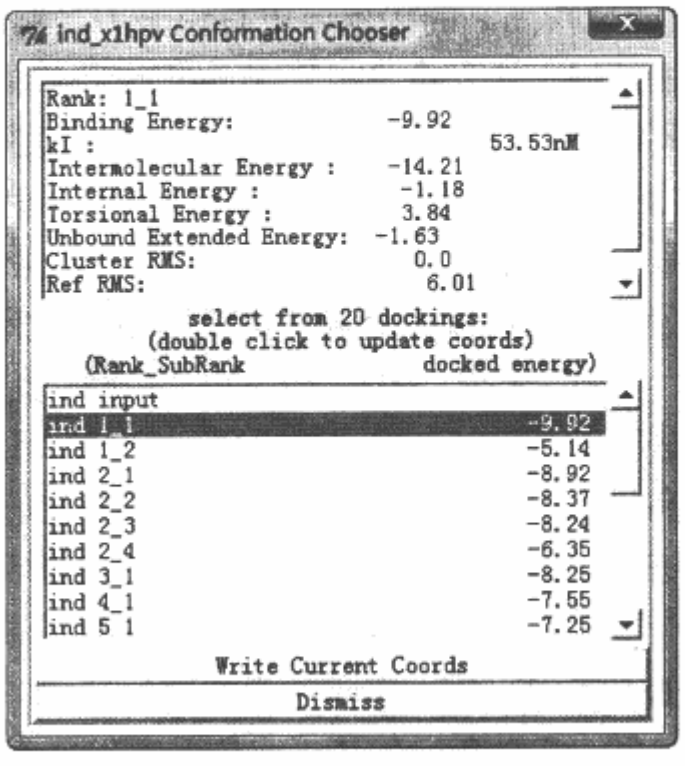


图 5-41　对接模式信息

③ 获得聚类信息　如上面提到的一样，聚类分析是评价对接结果优劣的主要指标之一。实际上，在默认情况下，聚类分析在完成对接运算的时候就已经被执行了。有兴趣的读者可以打开一个 DLG，并以 clustering 关键字进行搜索，然后观察 DLG 中给出的聚类信息。在 ADT 中查看聚类信息的方法为 Analyze | Clustering | Show，在执行完该操作后，程序会弹出一个包含了聚类信息的直方图（图 5-42）。通过观察该图，不仅可以清楚地获得聚类信息，而且还可以通过用鼠标点击柱状图中任一聚类来与 3D 窗口进行互动。如图 5-43 所示，在点击了第一类之后，在 3D 窗口中会出现一个类似于电影放映控制条的工具，用它可以将选定类中所有构象依次显示在 3D 窗口中。在这里需要提醒大家注意的是，之前没有执行过 Analyze | Conformations | Play 或 Play ranked by energy，则在点击聚类直方图时会出错。读者只需在点击聚类直方图之前运行一下 Analyze | Conformations | Play 即可避免此类问题。

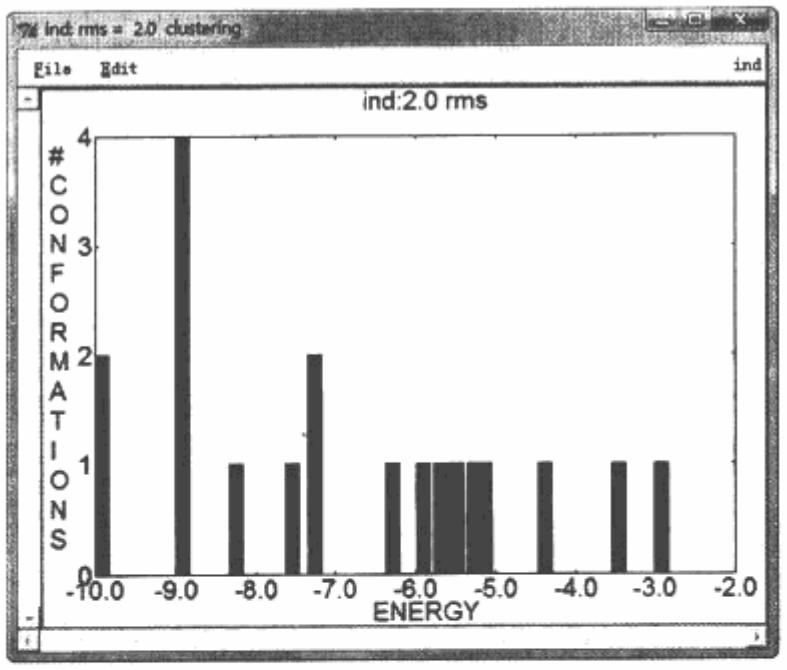


图 5-42　聚类信息图

在 AutoDock4 中，聚类分析的阈值默认为 2Å。对于一般的小分子配体来说，这个阈值是比较合适的，但是对于较大或较小，抑或是柔性键较多的分子来说，这个值就需要调整了，否则可能会得到非常不合理的聚类信息。一般来说，对于较大的或柔性键较多的配

体，需要将聚类阈值设置大些，比如 2.5～3 Å；而对于较小的配体，一般要将聚类阈值设置小一点，比如 1～1.5 Å。聚类分析其实在运行 AutoDock 的时候就已经被执行了，那么是否意味着如果想要得到不同阈值下的聚类信息就必须进行多次对接计算呢？这显然是不必要的。只需要通过外部工具对 DLG 文件中所提供的数据在不同的阈值下重新进行聚类分析就可以得到新的聚类信息，而 ADT 提供了这样一种工具。

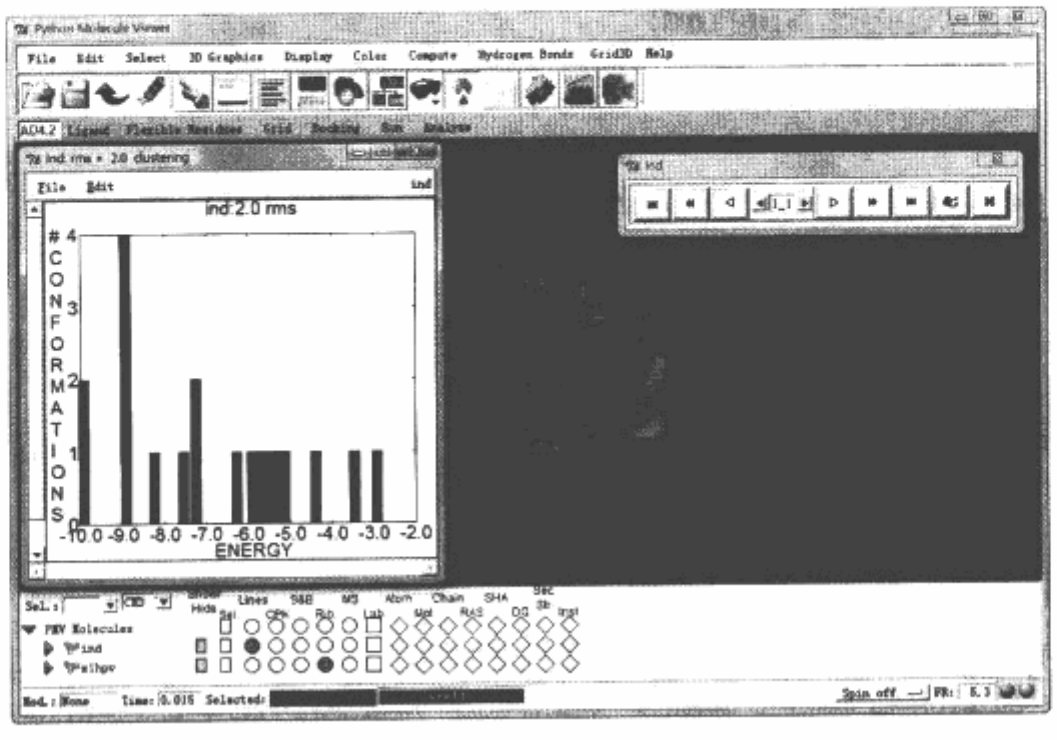


图 5-43　聚类来与 3D 窗口互动图

读者可以通过单击 Analyze | Clustering | Recluster 调出 Cluster Conformations 窗体，然后在 Tolerances 后面的文本框里填入新的聚类阈值（可以填入多个，具体格式见图 5-44），点击 OK 以进行新的聚类分析。之后，读者可以再次运行通过 Analyze | Clustering | Show 来显示聚类直方图。

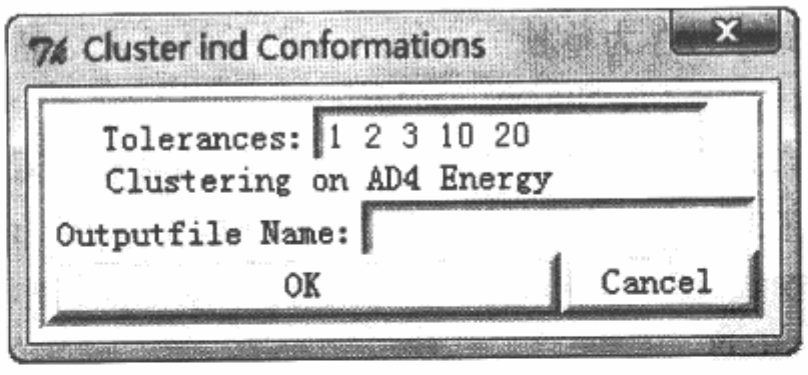


图 5-44　聚类阀值格式

④ 检查结合模式　在通过考察结合自由能及聚类信息之后，一般就已经选定了一些可能的结合模式。接下来是对选定的结合模式进行更进一步的分析，如氢键、疏水、静电等相互作用是否合理。在 ADT 中，可以把受体和选定的结合模式一起读入到 3D 窗口中，然后通过各种操作来检查选定的结合模式。由于这些分析涉及较多的背景知识，且 ADT 在分析功能上并不出众，因此不再进行详细的说明，有兴趣的读者可以自行研究。

（3）提取对接构象　读者可以利用 ADT 菜单栏上的 File | Save 方便地储存当前 3D 窗口中的分子，也可以通过写 shell 或 perl 脚本的方法从 DLG 文件中提取对接结果。例如，

读者可以利用如下 shell 脚本来提取对接得到的所有复合物构象：

```
###### Begin the Shell Script ######
ligand=ligandname
receptor=receptorname
write_all_complexes.py -d ./${ligand}_${receptor}.dlg -r ./${receptor}.pdbqt
-o ./pdbqts/${receptor}_${ligand}
for ((i=0;i<100;i++))
do
pdbqt_to_pdb.py -f ./pdbqts/${receptor}_${ligand}_${i}.pdbqt
-o ./pdbs/${receptor}_${ligand}_${i}.pdb
done
###### End the Shell Script ######
```

需要注意的是，在运行该脚本前，读者需要在当前工作目录下（DLG 文件所在目录）下分别建立名为 pdbqts 和 pdbs 两个文件夹。

三、实例分析

前面详尽阐述了用 ADT 和命令行来操作 AutoDock4 进行分子对接的步骤及方法，以下将通过两个完整的实例来向读者阐述 AutoDock 在实际科研工作中的应用。

1．半柔性对接：HIV-1 蛋白酶与其抑制剂的结合模式研究

为了验证 AutoDock4 的对接效果，在这个例子中选取晶体结构已知的 HIV-1 蛋白酶-抑制剂复合物来进行说明，具体操作步骤如下。

① 从 RCSB Protein Data Bank (http://www.rcsb.org/)下载蛋白酶-抑制剂复合物，PDB 号：1TCX。

② 将该复合物读入 VMD(http://www.ks.uiuc.edu/Research/vmd/) 或其他可视化分子模拟软件，并将受体、配体另存为单独的文件（注意不要包含结晶水或其他杂质），得到 ligand_1TCX.pdb 和 receptor_1TCX.pdb。

③ 用 ADT 或 python 脚本对 ligand_1TCX.pdb 和 receptor_1TCX.pdb 进行处理，生成配体和受体的 PDBQT 文件，分别得到 ligand_1TCX.pdbqt 和 grid_receptor_1TCX.pdbqt 和 flex_receptor_1TCX.pdbqt。在设置配体柔性键时选择所有可自由旋转的单键；在设置受体柔性残基的时候选择 ARG8。

注意：A、B 链上的都要选择。

④ 用 ADT 或 python 脚本对 ligand_1TCX.pdbqt 和 receptor_1TCX.pdbqt 进行处理，生成 GPF 和 DPF 文件，分别得到 ligand_1TCX.GPF 和 receptor_1TCX.DPF。

注意：生成 GPF 时所选择的受体一定要是选择受体的刚性部分，即 grid_receptor_1TCX.pdbqt。

⑤ 运行 AutoGrid4，生成下一步所需的 map 文件，主要参数设置为：npt=60 60 66, spacing=0.375, gridcenter=晶体结构中配体的质心。

⑥ 运行 AutoDock4，得到结合模式，主要参数设置为：ga_num_evals=2500000，ga_run=100。

结果：一共生成了 100 个对接模式，其中最好的结合模式与晶体结构的 RMSD 仅为

0.71Å，说明 AutoDock4 能够较好地预测出受体-小分子配体的复合物结构。

2. 多构象对接：HIV-1 整合酶与香豆素（coumarin）类抑制剂 NSC158393 的结合模式预测

在上面的例子中，为了更好地模拟受体-配体间的相互作用，利用了 AutoDock4 的新特性——半柔性对接。半柔性对接考虑了受体的侧链柔性，因此在一定程度上解决了在以往的全刚性对接中没有考虑受体活性口袋构象变化的问题。比起传统全刚性对接，AutoDock4 的半柔性对接已经取得了长足的进步，但是对于主链的柔性，AutoDock4 依然无能为力。为了在对接中考虑受体主链柔性，可以采用多构象对接（relaxed-model）的方法。所谓多构象对接，就是在进行分子对接之前，先通过分子动力学（MD）模拟等方法获得 N 个受体构象，然后将欲研究的配体分别与这 N 个受体构象进行对接，从而得到 $N\times$ga_run 个复合物构象。

在本例中，采用多构象分子对接的方法对 HIV-1 整合酶与香豆素（coumarin）类抑制剂 NSC158393 的结合模式进行了预测，具体步骤如下：① 从 RCSB Protein Data Bank 下载蛋白酶-抑制剂复合物，PDB 号：1BL3；② 用 Sybyl 中的 Loop Search 工具将缺失的残基补全；③ 用 NAMD 或其他动力学程序对经过处理的整合酶结构进行分子动力学模拟；④ 从得到的分子动力学轨迹中挑选出 12 个差异较大的构象作为分子对接中的受体结构；⑤ 用 ChemDraw 和 Chem3D 构建并优化 NSC158393 的立体结构；⑥ 用 AutoDock4 对 NSC15393 和通过 MD 模拟生成的 12 个受体结构进行分子对接，操作步骤和实例 1 中的步骤一致。

结果：通过多构象分子对接，共获得了 1200 个可能的 NSC158393 与 HIV-1 整合酶的结合模式。通过综合考虑这些结合模式的能量、聚类以及化学合理性，最终得到了一个较为合理的结合模式。通过这个实例，说明了多构象对接能够较好地考虑受体主链柔性。当受体活性口袋区域的柔性较大时，读者可以采用该方法来解决受体主链的柔性问题。

第六章

蛋白质-蛋白质复合物的结构模拟

第一节　蛋白质与蛋白质对接的意义

生物体内各种生命信息由不同的基因经转录、翻译传递到相应的蛋白质上并使其具有各自的生化特性及生物学活性，但每个蛋白质并不是独立地在细胞中完成被赋予的功能。细胞内的各种重要生理过程，如基因的复制、转录、翻译以及细胞信号转导、细胞周期调控、免疫反应等都是通过蛋白质-蛋白质分子之间的相互作用来进行的。各种蛋白质在细胞中通常与其他蛋白质相互作用形成大的复合体，在特定的时间和空间内完成特定的功能，而且有些蛋白质的功能只有在复合体形成后才能发挥出来。因此要充分理解和认识细胞复杂的生理过程，必须要解析蛋白质复合物结构并了解蛋白质相互作用过程发生的构象变化。此外，蛋白质复合物的结构可以为蛋白质界面提供信息，可以协助进行药物设计，可以帮助人们了解信号通路和评价复合物的亲和性。

由于当前研究结构的实验方法如 X 射线衍射和核磁共振技术的限制，解析蛋白质复合物的结构非常困难，造成了目前已知结构的蛋白质复合物非常有限。随着计算机处理能力的不断增强，理论模拟方法得到迅速发展和广泛应用，蛋白分子对接是一种预测蛋白质相互识别以及相互作用的工具。所谓蛋白分子对接就是根据两个已知蛋白质分子的三维结构，利用理论模拟手段考察它们之间是否可以结合，并预测复合物的结合模式。采用分子对接方法研究蛋白质分子间相互作用与识别，进而预测复合物结构。蛋白质分子对接已成为研究蛋白质与蛋白质相互作用问题的有力的计算机模拟手段，并可为实验工作提供有益的理论指导。

目前已经有了几款比较流行的大分子对接软件如 ZDOCK、RosettaDock（RDOCK）、Hex 等。在这一章里，主要为大家介绍其中两种。

第二节　利用 ZDOCK 来研究蛋白质-蛋白质的对接

在 DS 中，可以使用 ZDOCK 来实现蛋白质的对接计算。ZDOCK 是一种基于快速傅里叶转化相关性技术的刚性蛋白对接算法。算法中快速傅里叶转化相关性技术被用来所搜蛋白-蛋白系统的平动和转动空间。RDOCK 是一种基于 CHARMm 的能量优化过程，它被用来优化 ZDOCK 所寻找到的蛋白-蛋白复合物的结合构型，并使用能量打分函数给这些结合构型打分。本实例中，将会使用 ZDOCK 来进行蛋白对接的实验，分析对接的结果，并从中选取一些结合构型，利用 RDOCK 进行优化。本实例的蛋白对接实验中，选用牛 β-胰岛素（PDB ID 2ptn）作为受体蛋白，另外一个胰岛素抑制剂 CMTI-I（PDB ID 2sta，I 链）作为配体蛋白。该牛 β-胰岛素及其抑制剂 CMIT-I 复合物的晶体结构为 lppe。

一、设定一次 ZDOCK 计算

1．在同一个 3D 窗口中打开受体蛋白和配体蛋白

在 File 中，找到 2ptn.pdb 文件，双击。DS 将在一个新的 3D 窗口中打开该蛋白。然后，找到 2sta-I.pdb 文件，把它拖拽到同一的 3D 窗口中（图 6-1）。

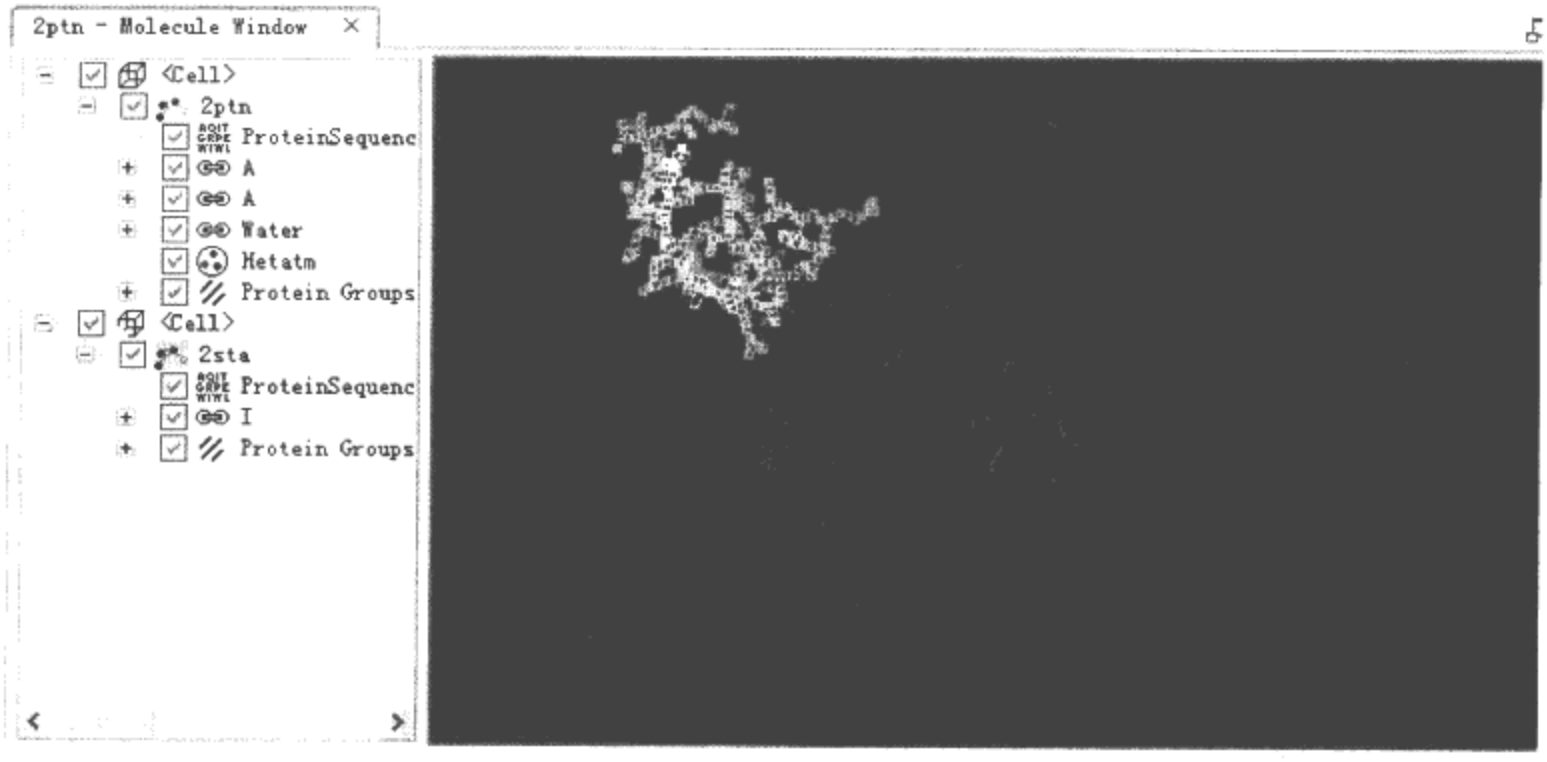


图 6-1　在同一个 3D 窗口中打开受体蛋白和配体蛋白

2．打开蛋白对接 Protocol（ZDOCK）并修改参数

Protocol 中，展开 Protein Modeling 文件夹，双击 Dock Proteins （ZDOCK），DS 将在 Parameter 中开打参数设置界面。

在 Parameter 界面中，点击 Input Receptor Protein 参数，选择 2ptn：2ptn；点击 Input Ligand Protein 参数，选择 2ptn：2sta；点击 Angular Step Size 参数，选择 15。

注意：ZDOCK 计算过程中可以采用两种欧拉角度进行结合构型的采样：6° 和 15° 。采样角度为 6° 时，预测结果更为准确，因为它的最终样本数包括 54000 个结合构型。尽管采样角度为 15° 时的结果准确性有所下降（因为它的采样数只有 3600 个），它的计算时间更短。

展开 Clustering 参数组，点击 RMSD Cutoff 参数，将该值设置为 6。点击 Interface Cutoff 参数，将该值设置为 9。点击 Maximum Number of Clusters 参数，将该值设置为 60。

本实例中使用的配体抑制剂较小，对于这样小的体系，将 RMSD Cutoff 设置为 6 Å 的 cluster 半径并同时将 Interface Cutoff 设置为 9 Å 能够产生更好的 cluster 结果。本实例中的 ZDOCK 运算，将不指定 blocked residues 也不会滤除结合位点的氨基酸。

如果有数据表明某些氨基酸残基不可能出现在蛋白-蛋白的作用界面，那么在受体蛋白中选中这些残基，然后设定 Receptor Blocked Residues 参数为 Create New Group from selection。同样的，也可以通过设置 Ligand Blocked Residues 参数来指定配体蛋白中不可能出现在界面上的残基。

如果有数据表明，某些氨基酸一定会出现在蛋白质-蛋白质作用界面上，那么选中它们。然后展开 Filter Pose 参数组，设置 Receptor Binding Site Residues 参数和 Ligand Binding Site Residues 参数为 Create New Group from selection。

Filter Poses 功能也可以在 ZDOCK 计算完后单独运行，使用 Process Poses（ZDOCK）protocol 即可。

3．运行 ZDOCK 并查看结果

在 Protocol 中，点击▶运行，等待结果完成。计算费时约 8min。计算结束时，DS 会出现 Job Completed 对话框。

在 Job 中，双击完成的计算任务。DS 将新打开一个 Report.htm 文件。

在 Output Files 部分，点击 ZDockResults.dsv 链接（图 6-2）。

二、分析 ZDOCK 结果

1．观察 ZDOCK 结合构型的聚类（Cluster）结果

在 Hierarchy View 中，展开 Docked Poses 和 Clusters。这将打开几组 Clusters 和 poses（图 6-3）。Cluster_1 为最大的聚类（包含的 poses 最多），后续的聚类越来越少。

图 6-2　查看 ZDOCK 结果

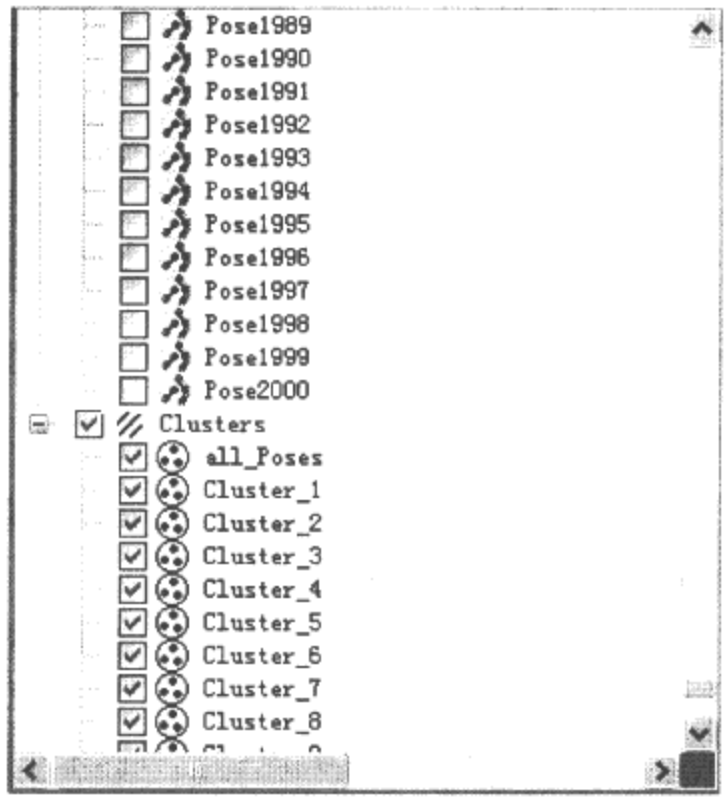


图 6-3　Clusters 和 poses

在 Data 中，点击 ProteinPose。

	sentative	ClusterSize	Density	lig_ctr_x	lig_ctr_y	lig_ctr_z	Rank	rec_ctr_x	rec_ctr_y	rec_ctr_z	ZDock Score	ZRank Score
1		8	8	17.257	23.783	43.124	1	-1.464	17.473	22.992	12.08	-83.595
2		1	2	15.874	24.986	43.958	2	1.991	17.973	19.752	8.54	-78.999
3		1	4	14.848	22.107	39.476	3	-1.464	17.473	22.992	10.08	-77.903
4		1	2	12.464	24.22	40.029	4	16.633	-0.036	27.382	7.88	-71.268
5		1	5	12.464	24.22	40.029	5	13.776	0.208	28.656	7.62	-64.824
6		8	7	14.848	22.107	39.476	6	-1.464	17.473	22.992	7.62	-64.645
7		1	4	12.421	24.042	42.787	7	-2.955	13.801	40.863	8.2	-63.785

Molecule / Atom / AminoAcid / Residue / Group / ProteinSequence / ProteinPose / DockedProteinPoseAssembly / Cell / AminoAcidChain / Bond / Chain / SolidSurface / Assembly /

这将显示如下关于 ZDOCK 计算后的信息：

- ZDock Score——包含每个对接构型的 ZDOCK 分数（PSC 打分函数），Pose 1 ZDOCK 分数最高
- Cluster——指明每个对接构型所属的聚类组
- ClusterSize——报告每个聚类含有的对接构型的个数
- Density——报告聚类过程中邻近的对接构象数目

关于聚类算法的详述，可以参考 Discovery Studio Help 中的“Clustering and analysis of

docked protein poses”。

在 Toos 中，选择 Protein Modeling。这将打开 Protein Modeling 的工具面板。

在 Analyze Docked Proteins 工具面板中的 Clustering 工具组中，点击 Show Cluster Center，然后点击 Show Full Cluster。这使 DS 的显示在所有的对接构型和聚类中心间进行切换。

2．对结合构型作图

以下的步骤将解释如何使用不同的方法，借助 plots 和 Analyze Docked Proteins 工具来帮助观看对接构型，并选择一部分 ZDOCK 的对接构型进行进一步的 RDOCK 优化。

在 Hierarchy view 中，按住 SHIFT，选择 Cluster_1～Cluster_10。这将选中 Cluster1~10。

在 Data Table 中，点击 ProteinPose，从菜单栏中，选择 Chart|3D Point Plot。这将打开一个 Choose Plot Axes 的对话框。在对话框中，选择 ZDock Score 作为 X 轴，Cluster 作为 Y 轴，Density 作为 Z 轴，Color 轴选择 ZDock Score。点击 OK。DS 将产生一个 3D Point Plot（图 6-4）。

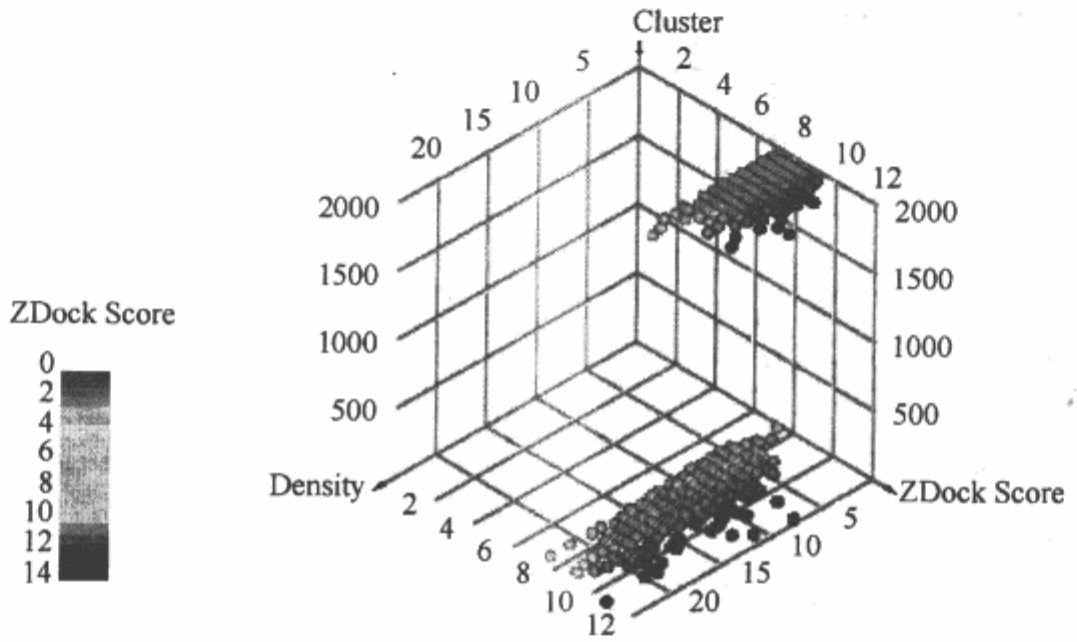


图 6-4　3D Point Plot

点击 ZDockResult-3D Point Plot，拖拽该窗口，以使能够同时看到 ZDockResults-3D 窗口（图 6-5）。

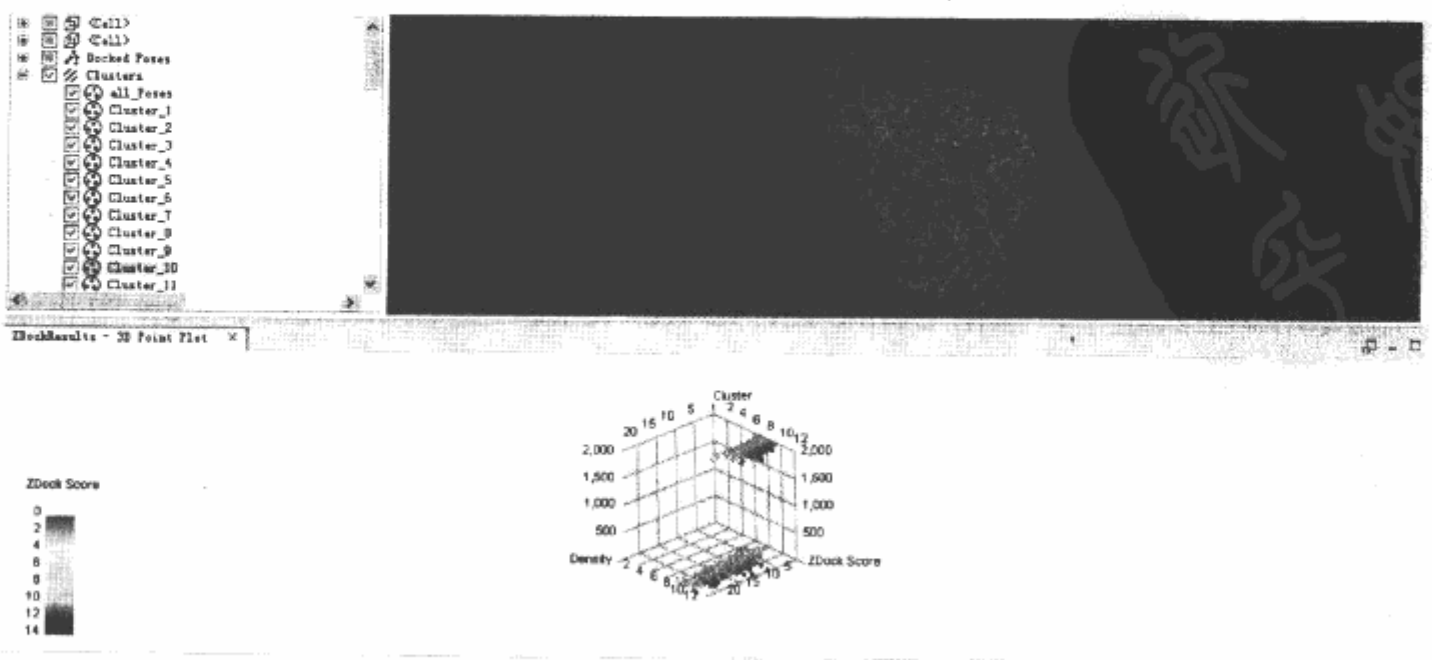


图 6-5　ZDockResults-3D 窗口

在 View toolbar 中，点击 Select 工具，选中 ZDockResult-3D Point Plot 窗口中的一些 high ZDock scores、high Density values 和 low Cluster number 的对接构型。选中的对接构型在 3D 窗口中也被选中。

点击 ZDockResult-3D 窗口，在 Hierarchy View 中的 Cluster 组中，点击 Cluster_2 将其选中。

在 Tools Explorere 中的 Analyze Docked Proteins 工具面板中，在 Display 工具组中，点击 Show Full Docked Complex。3D 窗口中将显示 Pose 329 的对接构型。Hierarchy View 中也会显示 Pose 329 蛋白复合物中受体和配体的名字：Pose329_2ptn 和 Pose329_2sta（图 6-6）。

在 Analyze Docked Proteins 工具面板中，点击 Show Next Docked Complex。3D 窗口中将出现 Pose 330 的对接构型。Hierarchy View 中也会显示 Pose 330 蛋白复合物中受体和配体的名字：Pose330_2ptn 和 Pose330_2sta（图 6-7）。

图 6-6　Pose 329 的对接构型

图 6-7　Pose 330 的对接构型

点击 Show Next Docked Complex 或 Show Previous Docked Complex，可以观察各个不同的对接构象。返回，并显示 Pose 1。在 Hierarchy View 中，点击 Pose1_2ptn。这将选中整个受体蛋白。

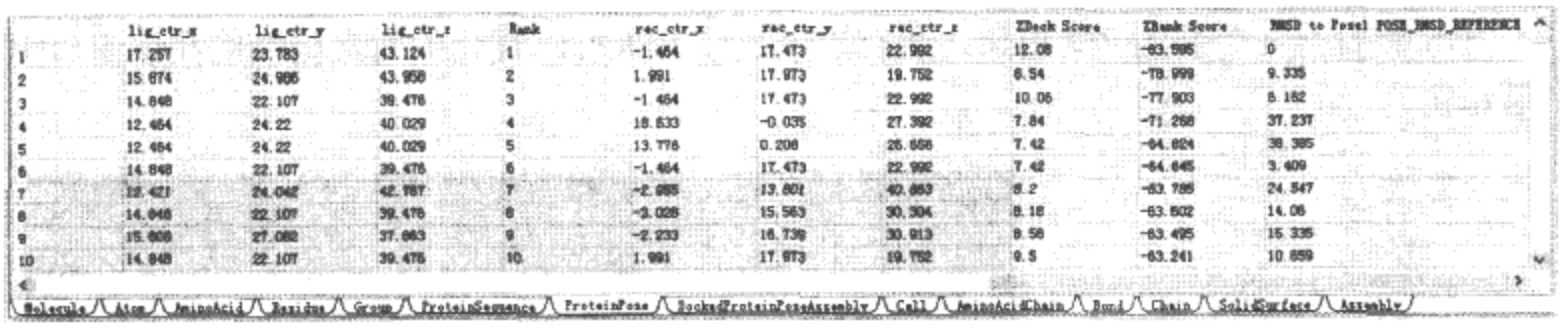
在 Analyze Docked Protein 工具面板中，点击 Identify Binding Interface。这将选择受体和配体对接界面的氨基酸。

在 RMSD 工具组中，点击 Define RMSD Reference。这在 Hierarchy View 中将添加一个叫 Pose 1 Pose_RMSD_REFERENCE 的组。这个 pose 将用作计算 RMSD 值的参考 pose。

3．计算 RMSD 值

在 Analyze Docked Proteins 工具面板中，RMSD 工具组，点击 Calculate Binding Site RMSD（ZDOCK）。

这将在 Data 中的 ProteinPose 添加一列名为 RMSD to Pose1 POSE_RMSD_REFERENCE 的列（图 6-8）。



	lig_ctr_x	lig_ctr_y	lig_ctr_z	Rank	rec_ctr_x	rec_ctr_y	rec_ctr_z	ZDock Score	ZRank Score	RMSD to Pose1 POSE_RMSD_REFERENCE
1	17.257	23.783	43.124	1	-1.464	17.473	22.992	12.08	-83.595	0
2	15.674	24.986	43.958	2	1.991	17.973	19.752	8.54	-78.999	9.335
3	14.848	22.107	39.476	3	-1.464	17.473	22.992	10.06	-77.903	8.162
4	12.464	24.22	40.029	4	18.533	-0.035	27.392	7.84	-71.268	37.237
5	12.464	24.22	40.029	5	13.776	0.206	25.656	7.42	-64.824	38.385
6	14.848	22.107	39.476	6	-1.464	17.473	22.992	7.42	-64.645	3.409
7	18.421	24.042	42.787	7	-2.685	13.801	40.853	6.2	-63.785	24.547
8	14.848	22.107	39.476	8	-3.028	15.563	30.304	8.18	-63.602	14.06
9	15.609	27.082	37.663	9	-2.233	16.738	30.913	9.58	-63.495	15.335
10	14.848	22.107	39.476	10	1.991	17.973	19.752	9.5	-63.241	10.859

Molecule / Atom / AminoAcid / Residue / Group / ProteinSequence / ProteinPose / DocksProteinPoseAssembly / Cell / AminoAcidChain / Bond / Chain / SolidSurface / Assembly

图 6-8　添加 POSE_RMSD_REFERENCE 的列

Cluster_48 中的许多 poses 都与 PDB 1ppe 中的结合模式相近。

提示：在 Data 中的 ProteinPose 中，双击 RMSD to Pose1 POSE_RMSD_REFERENCE 列，可以将 pose 排序（图 6-9）。

三、利用 RDOCK 优化对接构型

1．设置 RDOCK protocol

关闭 ZDockResults-3D Window 和 ZDockResults-3D Point Plot Window 窗口。点击 Dock Proteins (ZDOCK)-Html Window，在 Output Files 部分，点击 ZDock-Results.dsv 打开一个新的 ZDockResults 的 3D 窗口。

RDOCK 需要受体和配体都用 CHARMm Polar H 力场进行赋值。

	Cluster	RMSD to Pose1 POSE_RMSD_REFERENCE
1	48	0
2	48	1.96
3	48	2.191
4	48	2.354
5	48	2.763
6	48	3.409
7	48	4.417
8	48	4.546
9	2,001	6.162
10	2,001	7.938
11	2,001	8.119

图 6-9　将 pose 排序

Tools 中，选择 Simulation。这将打开 Simulation 工具面板。在 Forcefield 工具面板中，选择 CHARMm Polar H，然后点击 Apply Forcefield。这将把 CHARMm Polar H 力场赋予受体和配体蛋白。工具面板中将显示 ZDockResults typed with CHARMm Polar H 的状态。

在 Protocols 中的 Protein Modeling 文件夹，双击 Refine Docked Proteins (RDOCK) protocol。这将在参数中打开 RDOCK 的参数设置。

在 Protocols 中，点击 Input Typed Receptor Protein 参数，选择 ZDockResults:2ptn，点击 Input Typed Ligand Protein 参数，选择 ZDockResults:2sta_I。

本例中，将选用排名前六位的对接构型中的一些来进行 RDOCK 优化。考虑到晶体结构中，抑制剂蛋白 CMTI-I 的结合 loop 中 N 端的 Arg505 是主要的锚定氨基酸，观察前六位的对接构型，发现只有 poses1、poses2、poses3 和 poses6 满足上面的要求。所以将采用这 4 个 poses 来进行优化。

在 Hierarchy View 中的 Docked Poses 组中，按住 CTRL 键，点击选中 poses1、poses2、poses3 和 poses6。在 Parameter Explorer 中，点击 Input Poses 参数，选择 Create NewGroup from selection…系统默认生成一个新的名为 Group 的组。

在 Protocl 工具条中，点击 ▶ 运行，等待计算完成。计算费时为约几分钟。计算结束时，DS 会出现 Job Completed 的对话框。

2．观察结果

在 Jobs 中，双击 job。DS 将打开一个名为 Report.htm 的文件。点击该窗口 Output File 部分的 View Results。

DS 将打开一个新的 3D 窗口，所有四个经过 RDOCK 优化的对接构型均在该窗口中（图 6-10）。

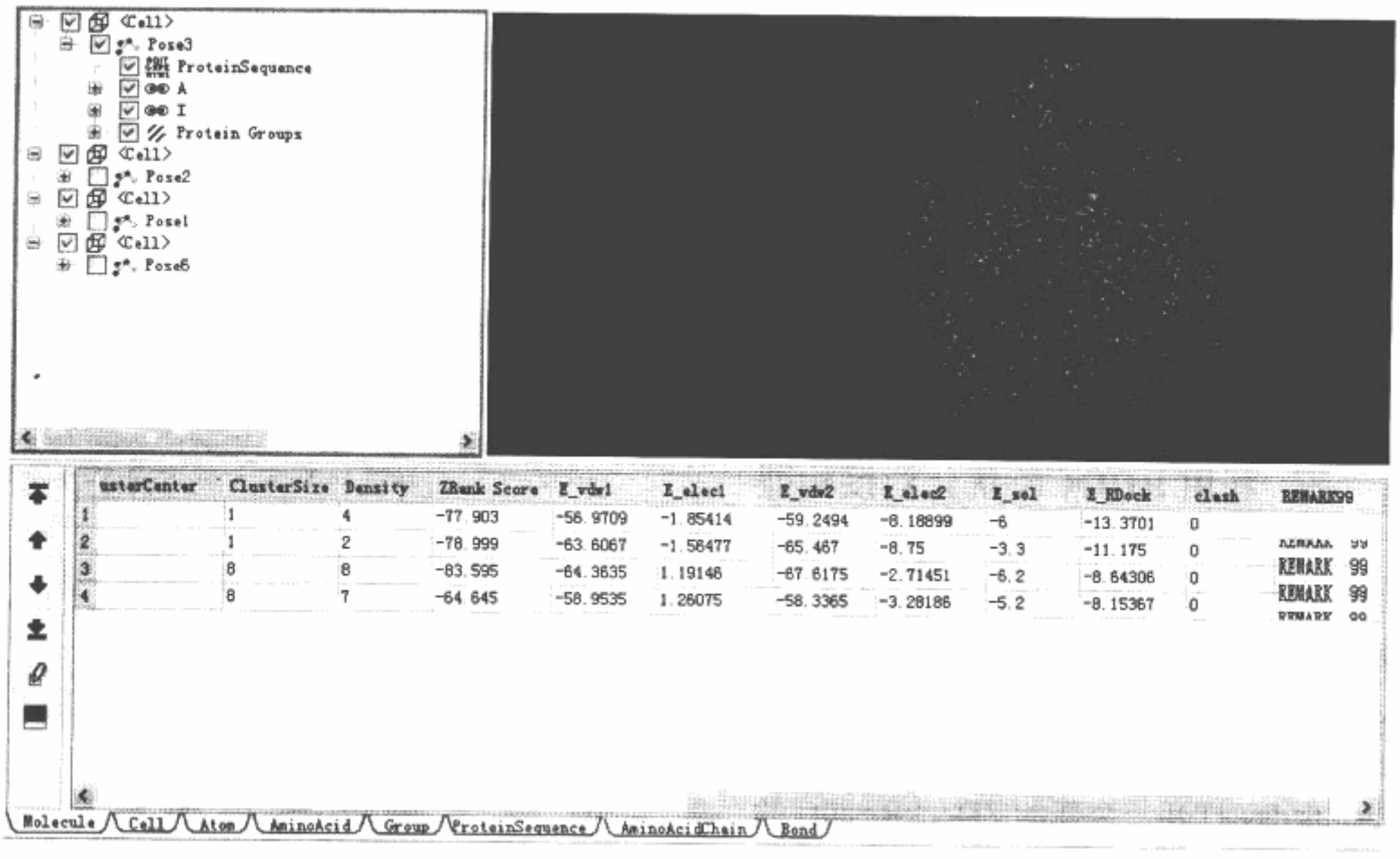


图 6-10　对接 3D 窗口

在表格中，有 RDOCK 所计算的各能量值的列表。

- E_elec1 and E_elec2——经过第一、二轮 CHARMm 能量优化后蛋白质复合物的静电能
- E_vdw1 and E_vdw 2——经过第一、二轮 CHARMm 能量优化后蛋白质复合物的范德华非键作用能
- E_sol——蛋白质复合物的去溶剂化能
- E_RDock——RDOCK 分数的定义为：E_elec2 + beta*E_sol

在表格中，点击⤒ 可以看到第一个 pose 的 3D 构象。分别点击 ⬆ 和 ⬇ 可以查看上一个和下一个 pose 的结构。

3．选择相互作用界面的氨基酸

在 Hierarchy View 中，展开 Selected pose，选择受体蛋白 A。

在 TooLs 中，选择 Protein Modeling。这将打开 Protein Modeling 工具面板。

在 Analyze Docked Proteins 工具面板中的 Display 工具组中，点击 Define Receptor。这将定义 A 为受体蛋白。

在 Analyze Docked Proteins 工具面板中的 Display 工具组中，点击 Identify Binding Interface。这将选中受体蛋白和配体抑制剂间相互作用的氨基酸。

提示：用以定义结合界面的 Distance Cutoff 值可以在 Preferences 中设定。从菜单栏中，选择 Edit | Preferrences | Protein Docking。默认值为 5Å。

四、使用 RMSD 分析结合界面的氨基酸

本小节中，将把对接的构型与晶体结构中的蛋白质复合物 PDB lppe 进行比较。

1．打开 PDB 1ppe

从菜单栏中，选择 File | Insert From | URL。这将打开一个从 URL 中 Insert 结构的对话框。输入 1ppe，点击 Open。这将在以后的 3D 窗口中插入 1ppe 的结构。

在 Hierarchy View 中，展开 Cell，然后展开 1ppe。选中 water，点击 Delete 键。这将删除 1ppe 中的水分子。在 Hierarchy View 中，展开 Cell，然后展开 1ppe。从菜单栏中，选择 Sequence | Show Sequence。这将打开一个 Sequence 窗口（图 6-11）。

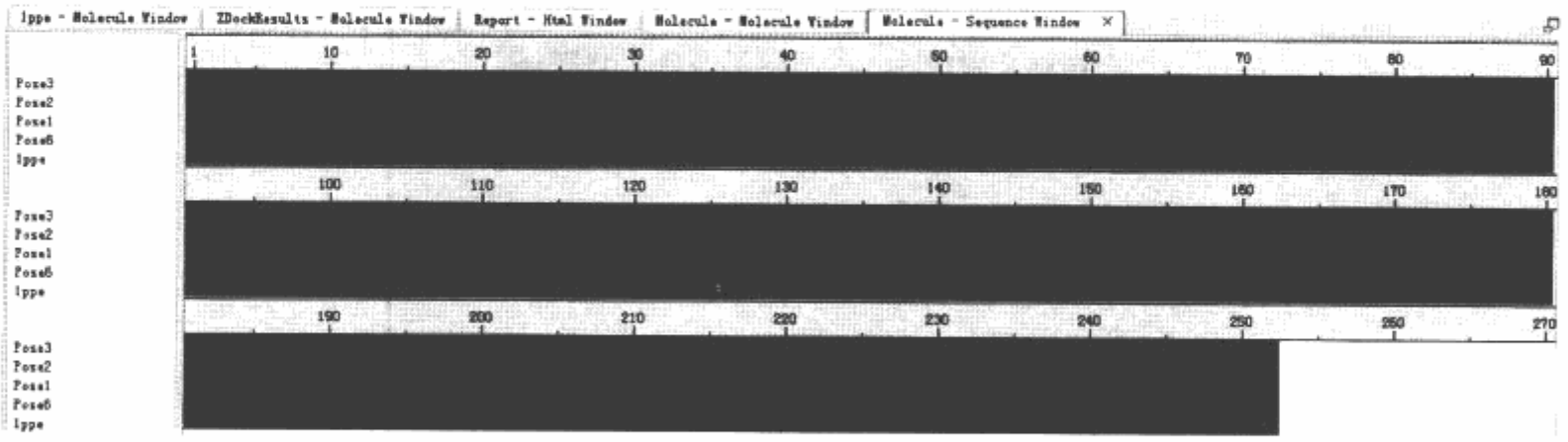


图 6-11　Sequence

在计算 RMSD 之前，对接的构型需要与 1ppe 基于受体蛋白结构进行叠合。

2．将对接构型与 1ppe 叠合

点击 Molecule 窗口使之激活。在 Hierarchy View 中，点击 chain E。这将选中 1ppe 中的受体蛋白。从菜单栏中，选择 Edit | Preferences。这将打开 Preferences 对话框。从菜单栏中，选择 Structure | Superimpose | By Sequence Alignment…这将打开 Sequence Alignment 对话框。1ppe 作为参考分子。

对话框（图 6-12）中，选取所有的 poses。点击 OK。

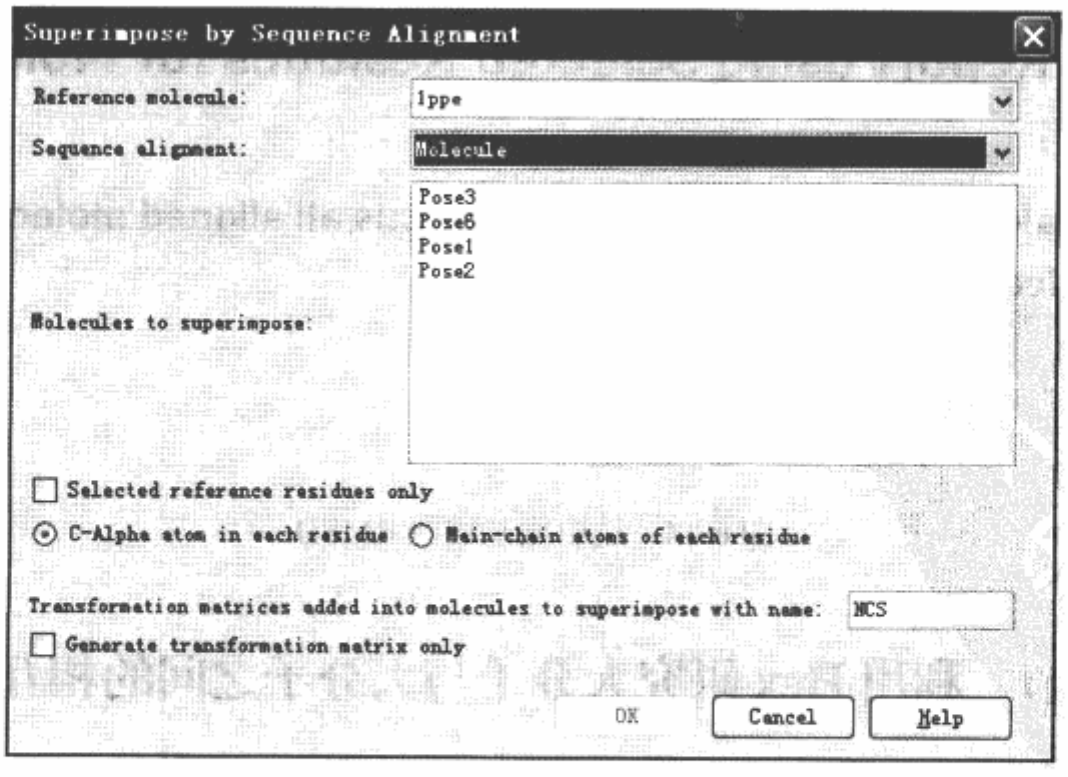


图 6-12　Sequence Alignment

在 Molecule 窗口中，核实 1ppe 受体蛋白 E 仍然被选中。

在 Tools 中的 Analyze Docked Proteins 工具面板中，先点击 Define Receptor，再点击

Indentify Binding Interface。这将选中受体-受体蛋白中的氨基酸。

3．计算结合界面氨基酸的 RMSD 值。

从菜单栏中，选择 Structure | RMSD | By Sequence Alignment…这将打开 RMSD by Sequence Alignment 的对话框（图 6-13）。

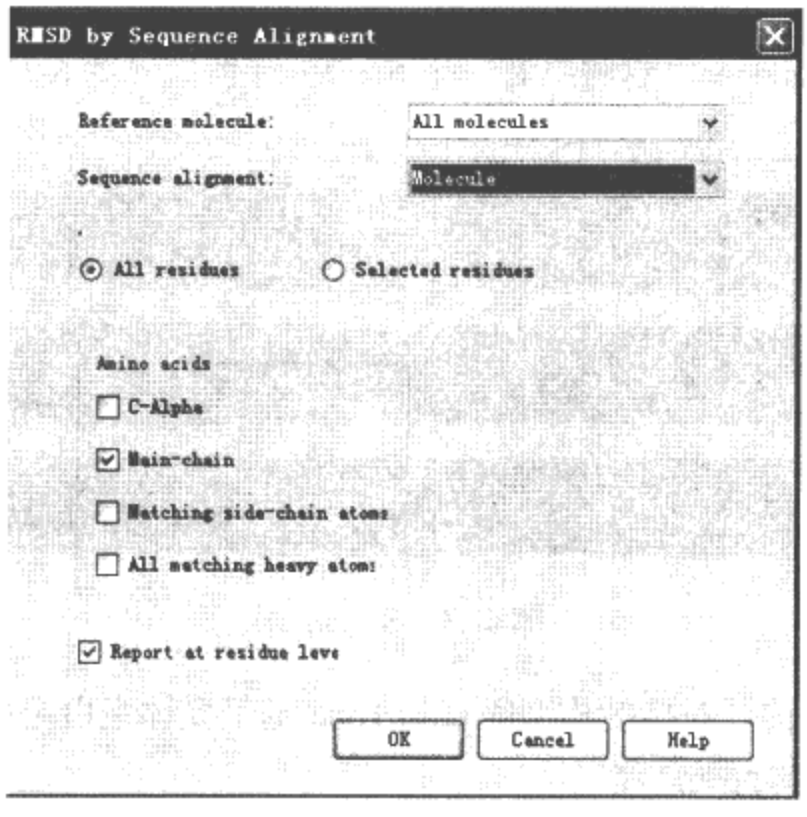


图 6-13　RMSD by Sequence Alignment

在该对话框中，Reference molecule 选择 1ppe。点击 Selected residues，关闭 Report at residue level 选项。点击 OK。DS 将在一个新的窗口中打开一个基于选中氨基酸的 RMSD 报告（图 6-14）。

RMSD Report using Selected Residues for Molecule

RMSD Calculation of reference 1ppe versus all aligned molecules

Molecule	Reference	Main-chain
Pose3	1ppe	2.228
Pose2	1ppe	1.809
Pose1	1ppe	3.436
Pose6	1ppe	4.398

图 6-14　选中氨基酸的 RMSD

第三节　利用 Hex 研究大分子与大分子之间的相互作用

上一节介绍了比较流行的大分子对接软件 ZDOCK，在本节，主要介绍用 Hex 研究大分子与大分子之间的相互作用。

Hex 是由苏格兰 Dave Ritchie 小组开发的一种研究大分子与大分子之间相互作用的软件，它不仅可以实现蛋白质-蛋白质的对接，还可以实现蛋白质-核酸的对接。与其他单分

子对接软件相比，Hex 具有简单易用和运算速度快两大优势。所谓简单易用，是因为 Hex 是一款图形用户界面化的对接软件。在 Hex 中，从对接前分子的处理到对接后的分析，所有的操作都可以在图形用户界面下完成，这就使得初次接触大分子对接软件的用户能够很快的熟悉 Hex 的操作。此外，由于 Hex 在对接中采用了 Spherical Polar Fourier 算法，因而使得其运算速度大幅提高。Hex 的设计者甚至宣称其计算速度比采用传统快速傅里叶变换（fast Fourier transform）算法的对接程序运算速度快 10 倍。

一、Hex 的安装

Hex 的安装很简单，只需要到 Hex 的官方网站(http://www.csd.abdn.ac.uk/hex)下载 Hex，然后在 Windows 下进行安装就可。安装完成后启动 Hex，可以看到 Hex 主界面（图 6-15）。

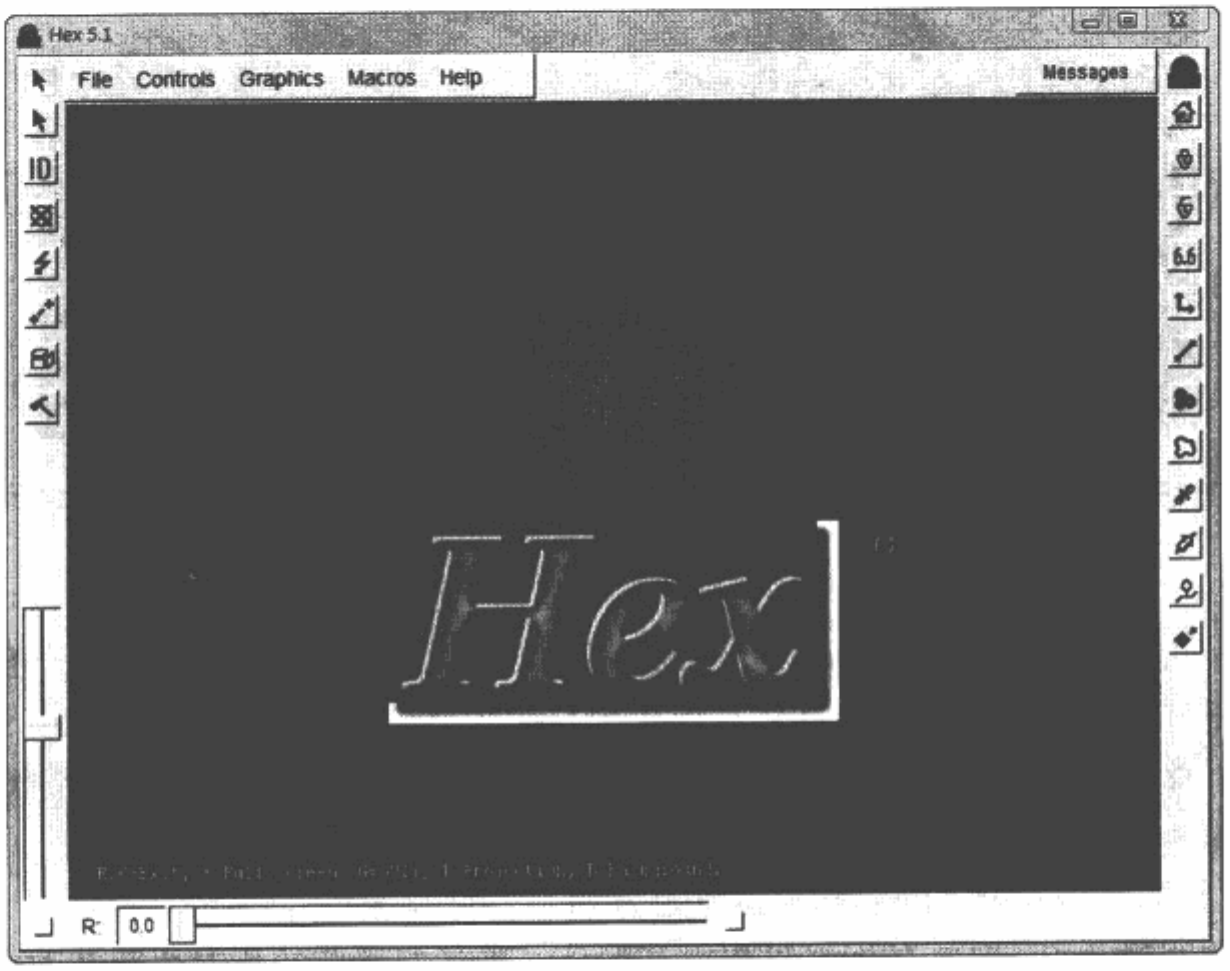


图 6-15　Hex 主界面

二、Hex 的基本应用

Hex 的操作比较简单，对接中的所有步骤都可以在 Hex 所提供的图形用户界面下完成。Hex 的操作流程如图 6-16 所示，可以分为 4 个步骤。

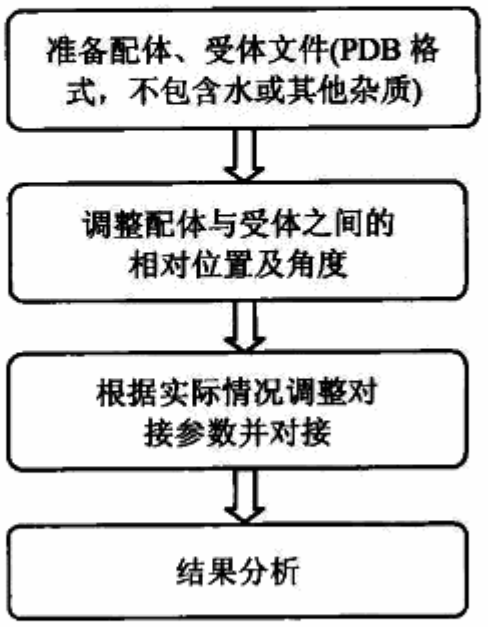


图 6-16　Hex 的操作流程

1．配体和受体文件的准备

Hex 的输入文件只需要两个 PDB 文件（一个配体，一个受体）。在进行对接之前，必须确保这两个 PDB 文件不能包含结晶水或其他杂质的数据。读者可通过 File | Open | Receptor 和 Ligand 分别读入受体及配体文件。对于 PDB 文件中所含的杂质原子，读者可以用 Hex 中 Controls | Heteros 菜单调出的 Hetero Control 窗体（图 6-17）来将其屏蔽掉（取消选择 Enable Solvent 及 Enable Hetero 复选

框），被屏蔽的杂质原子在 3D 窗口中用灰色表示。

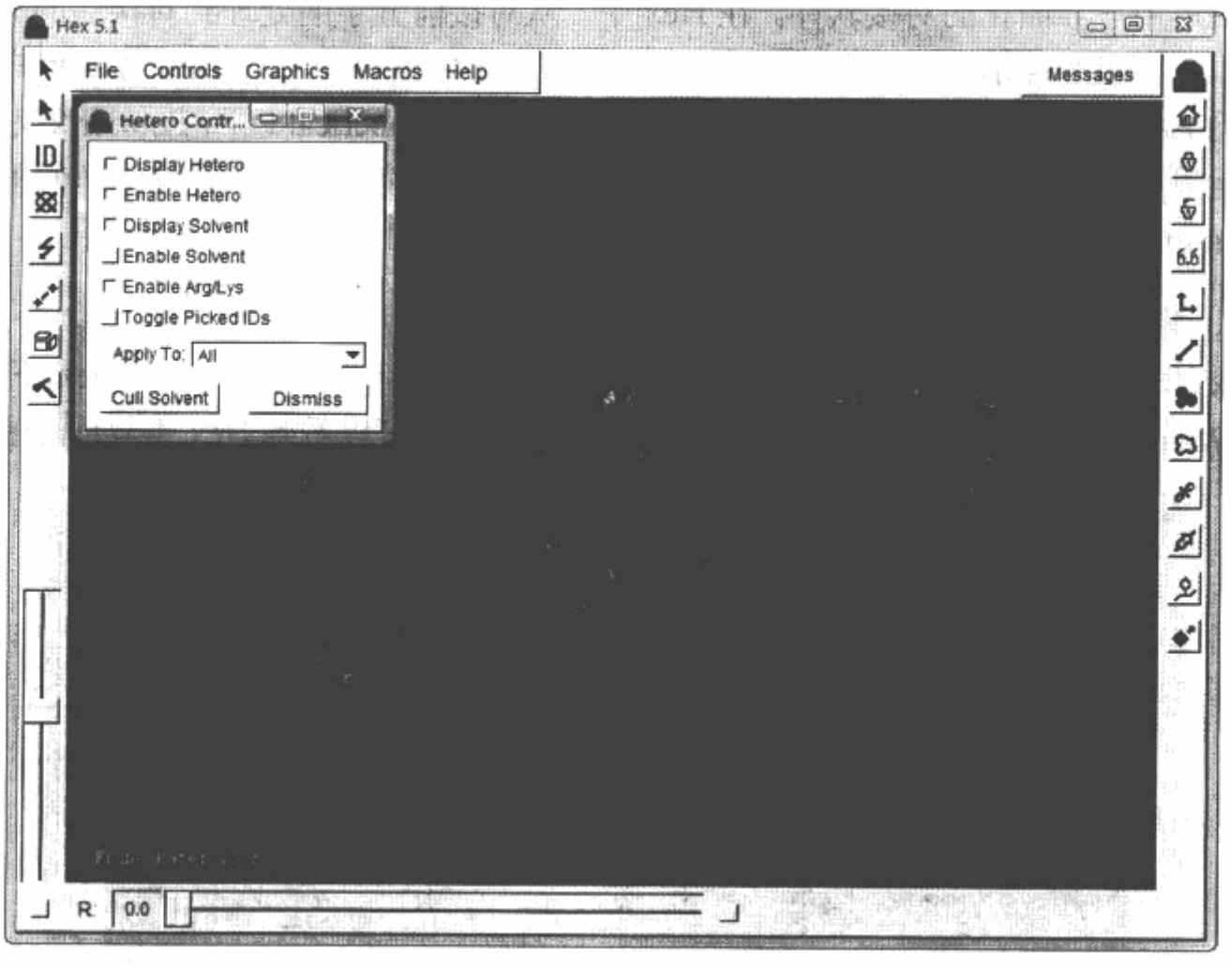


图 6-17　配体和受体文件的准备

此外，也可以通过 File | Open | Complex 选择性的读入一个复合物文件。若读入了复合物，则在对接过程中 Hex 会自动计算出每个对接结果与该复合物的 RMSD。

注意：必须保证该复合物的原子数、链及编号与受体及配体相一致，具体情况参考 Hex 5.1/examples/目录下的例子。

2．调整配体与受体之间的相对位置及方向

首先要明确一点，Hex 的对接结果是和配体、受体对接前的相对位置及方向相关的。如果在对接之前，能够通过某种途径得到欲研究的配体、受体之间相互作用的某些信息，那么就能够利用这些宝贵的信息来指导接下来的对接计算。比如可以通过查阅文献获得某些界面残基的信息，那么接下来，就可以在对接前调整配体、受体之间的相对位置及方向，然后在对接时限制空间搜索的方向。利用这种限制性的对接策略，可以较准确地找到正确的结合模式。

（1）调整配体、受体之间的距离　在 Hex 中，配体、受体之间的距离是通过测量配体、受体原点（默认情况下为重心）之间连线的长度来确定的。当这个长度大于 30Å 时，由于算法和所采用的势函数的原因，Hex 不能够给出可靠的对接结果。当 Hex 读入受体和配体之后，它会通过简单的计算粗略地猜测出配体、受体可能的结合模式。因此，在一般情况下，配体、受体之间的距离并不会大于 30Å。但是，当处理的受体、配体体积非常巨大的时候，往往这个距离都是大于 30Å 的，因此就有必要通过重新设置配体、受体各自的原点来减小二者之间的距离。设置配体、受体原点的方法如下。

首先，点击 Hex 主界面左侧工具栏上的 按钮以切换到“原点编辑模式”（Edit Molecule Origins mode），再点击右侧工具栏上 按钮显示当前配体、受体原点间的连线。

然后，通过鼠标左键点击欲设置为原点的位置，然后点击左侧工具栏上的按钮重新将鼠标模式切换为普通模式。

最后，通过右侧工具栏的按钮显示新设置的原点之间连线，可以发现，此时的连线已经和最初的大不一样了。

注意：现在设置的原点仅仅对当前的操作有效。当存储编辑过的配体或受体的结构时，原点信息并不会随之保存。因此，每一次用 Hex 做对接之前都要检查一下分子之间的距离。

（2）调整配体、受体之间的相对方向　正如上面提到的，在做对接之前，往往希望能获得各种先验的信息，从而对接下来的分子对接进行指导。当通过各种途径获得了这些宝贵的信息之后，就可以通过限制性对接降低假阳性（false positive）的概率，从而更准确地获得合理的对接模式。设置配体、受体相对方向的方法如下。

首先，点击 Hex 主界面左侧工具栏上的按钮以切换到所谓的“方向编辑模式”（Edit Molecule Orientations mode），再点击右侧工具栏上按钮显示当前配体、受体原点间的连线。

然后，用鼠标左键双击配体或受体。此时，被点击的分子的应该变成了虚线，这说明该分子已经被选中。

最后，可以通过鼠标左键拖拽选中的分子以改变分子间距离（注意，在该模式下并不能改变分子原点），也可以通过鼠标右键拖拽选中的分子以改变配体、受体的相对方向。

除了用上面提到的方法设置分子原点、距离及相对方向以外，也可以通过 Controls | Orientations 打开 Orientation Control 面板（图 6-18）对这些参数进行精确控制。

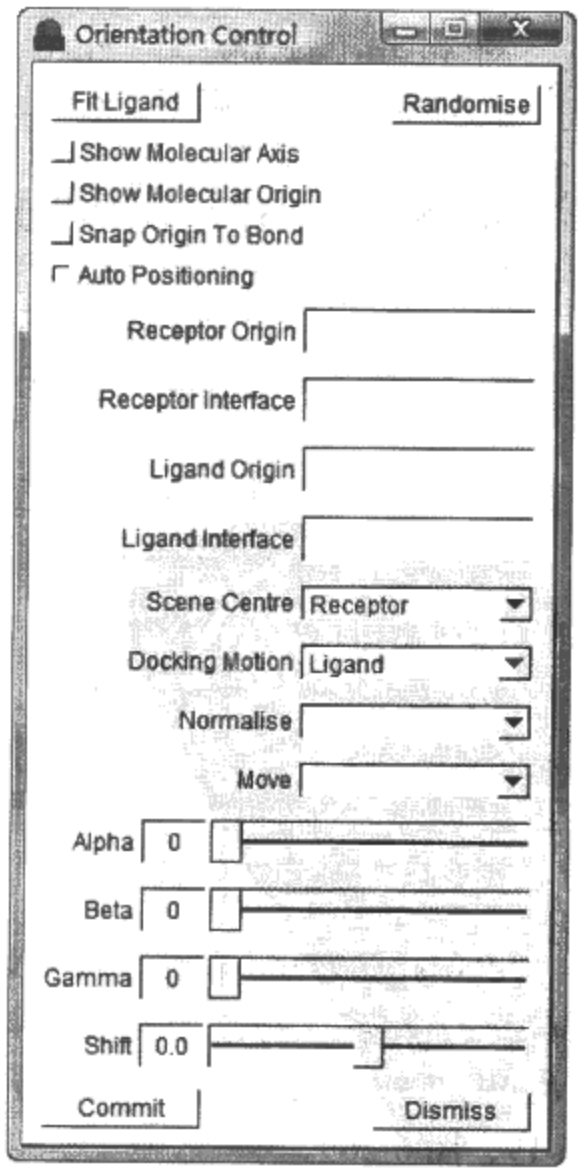


图 6-18

（3）对接　在设置好分子原点、分子间距离和相对方向之后，就可以用 Hex 进行对接了。在 Hex 中，对接中用到的所有参数都可以通过 Controls | Docking 调出的 Docking Control 面板进行设置（图 6-19）。首先说一下 Docking Control 面板中各参数的意义。

Steric Scan 和 Final Search：粗糙搜索和精细搜索，搜索精度默认值分别为 16 和 20。在 Hex 中，对接共分两步进行。第一步为较粗糙的搜索，即 Steric Scan；第二步为较精细的搜索，即 Final Search。当完成第一步搜索后，Hex 会自动放弃那些打分较差的结果，只有剩下结果才进入第二步计算。在 Hex 5.0 中，虽然二者的默认精度分别为 16 和 20，但是 Hex 的设计者建议使用的精度为 20/25 或 20/30。当受体或配体柔性较大时，建议使用 20/25；当柔性较小时，采用 20/30。在 Hex 5.1 中，经过大量的测试，这两步搜索的计算精度的默认值已经被调整为 18 和 25。因此对于一般的对接，Hex 5.1 的默认搜索精度已经足够了。此外，若取消 Steric Scan 复选框，则 Steric Scan 的精度由 Final Search 的精度共同控制。

Correlation Type：对接类型，有两种类型可供选择，即 Shape Only（默认）和 Shape+Electrostatics。当

选择 Shape Only 时，在对接中 Hex 只考虑受体、配体之间形状的契合而不考虑静电贡献；当选择 Shape+Electrostatics 时，Hex 会综合考虑形状和静电的贡献。值得一提的是，Hex 并不提倡在所有的对接中都采用 Shape+Electrostatics 模式，因为 Hex 中的原子电荷库还不够完善。但是，对于蛋白与核酸的对接，由于核酸分子带有强烈的负电荷，因此在对接时必须要考虑静电的贡献。Hex 只在 Final Search 时计算静电贡献。

FFT Mode：算法类型，有 1D Classic Hex、3D Fast Lite（默认）及 5D Mega Memory 三种模式可供选择。就计算速度来说，5D > 3D > 1D，但是搜索算法越复杂则所需的内存越多，读者可根据自己计算机的性能来选择算法。

Grid Dimension：用于计算 Surface skin expansion coefficient 的格点大小，可选择 0.4～1.0 Å。这个参数一般不需要修改，选择默认的 0.6 Å 即可。

Receptor Range：受体旋转（rotation）搜索范围，可选择 0～180°，默认为 180°。该角度越大则代表在受体上的搜索范围越大。当设置 Receptor Range = N° 时，则受体可相对于分子原点间连线（Intermolecular Axis）的旋转角度为 $\pm N/2$°（图 6-19）。需要注意一点的是，在对接中，受体实际上是不做任何运动的。受体所有的运动实际上都被转化为了配体的运动。

Ligand Range：配体旋转搜索范围，可选择 0～180°，默认为 180°。该角度越大则代表在配体上的搜索范围越大。当设置 Ligand Range = N° 时，则配体可相对于分子原点间连线（Intermolecular Axis）的旋转角度为 $\pm N/2$°（图 6-19、图 6-20）。

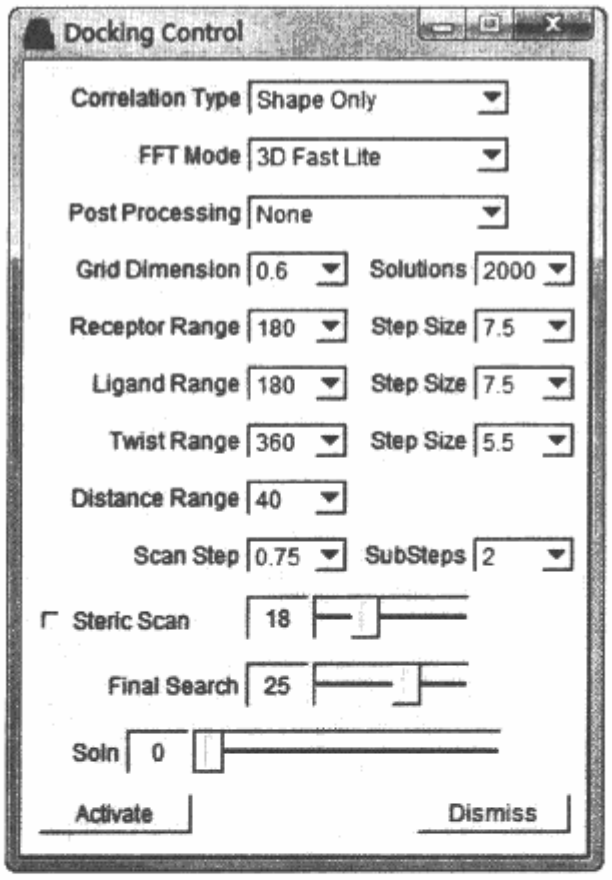


图 6-19　Docking Control 面板

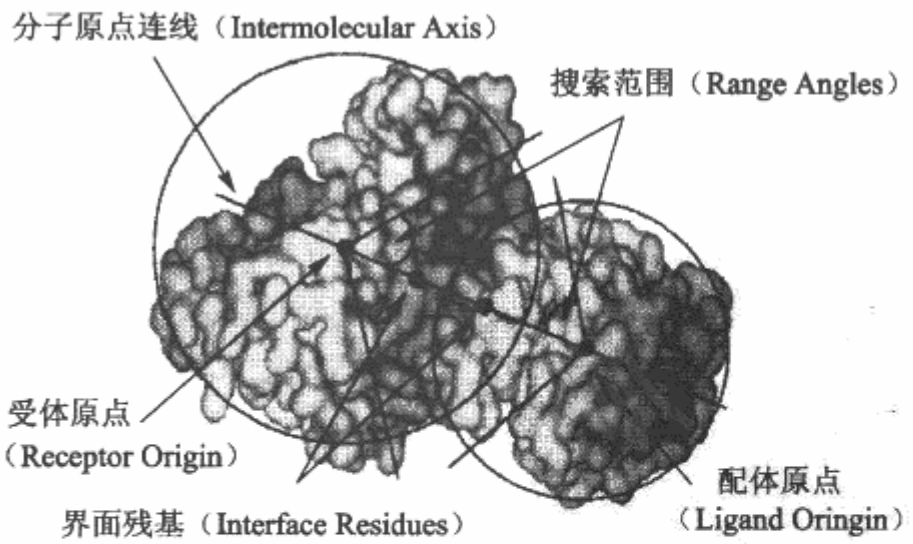


图 6-20　搜索范围限制示意图

Twist Range：配体相对于受体的扭转（twist）搜索范围，可选择 0～360°，默认为 360°。该角度越大则代表在配体可相对于受体的扭转角度越大。当设置 Twist Range = N° 时，则说明配体可相对于分子原点间连线（Intermolecular Axis）的扭转角度为 N°（图 6-19、图 6-20）。

Step Size：旋转/扭转搜索的步长，对于旋转搜索默认步长为 7.5°，而对于扭转搜索

默认步长为 5.5°。

Distance Range：距离搜索范围，可选择 1～60 Å，默认值为 40 Å。当设置 Distance Range = *N* 时，其距离搜索范围为 ±(*N*+1)/2 Å（1 Å 表示初始状态）。

Scan Step：Steric Scan 中所采用的距离搜索步长，默认为 0.75 Å。

SubSteps：Final Search 中所采用距离搜索的步长权重的倒数，默认为 2。当设置 Scan Step = *N* Å，且设置 SubSteps = *M* 时，Final Search 中所采用的距离搜索步长为 *N*/*M* Å。

Solutions：最终保留的对接结果数目，默认为 2000 个。

Post Processing：对接后的处理，有 None（默认）、Bumps+Volumes、MM Energies 及 MM Minimisation 四种模式可选。选择 None 表示在对接之后不做任何处理；选择 Bumps+Volumes 则 Hex 在对接结束后会计算每一种结合模式中的配体与受体之间的空间冲突，即所谓的 Bumps，该数据可用于聚类分析（设置 Clustering Control 面板上的 Bump Threshold 参数）；选择 MM Energies 或 MM Minimisation 则在对接结束后，Hex 会计算 Rigid Body 能（对于 MM Energies）或用分子力学进行能量优化（对于 MM Minimisation）。这里需要注意的一点是，当对接前后配体、受体之间相对位置变化很大时最好不要使用 MM Energies/ Minimisation，因为这两部分的程序还不完善。

Soln：对接结果显示滑动条。

当设置好所有的对接参数后，可以点击 Docking Control 面板最下方的 Activate 按钮进行对接，此时会弹出一个对接进程对话框（Hex Progress）来显示当前的对接进程（图 6-21）。当 Hex 完成所有运算后会将对接结果自动显示在 3D 窗口中，读者可以利用键盘上的 Page up/down 及 home/end 来浏览对接结果。在这里需要提醒大家注意的一点是，Hex 的聚类是在对接过程中完成的，但在完成对接之后，如果用户修改了（Contorls | Clustering，图 6-22）聚类参数中的 Max Cluster、Cluster Window 或 RMS Threshold，Hex 都会对当前的对接结果进行重新聚类。此外，单独修改 Bumps Threshold、Sort Solutions by 和 Display Cluster 并不会激活重新聚类。

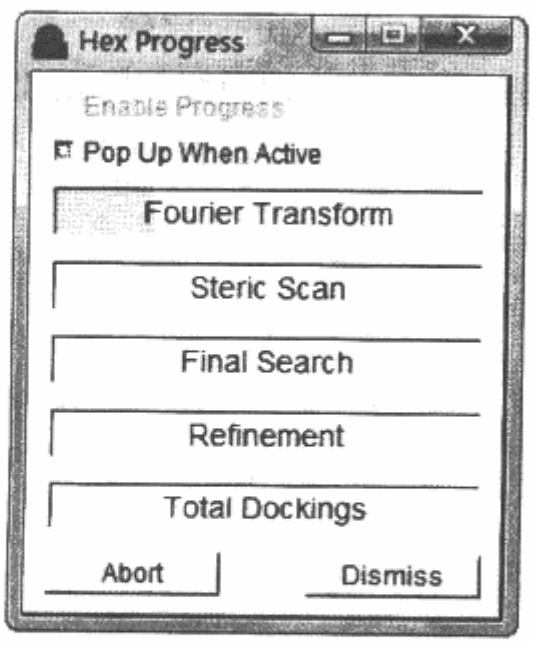


图 6-21　对接进程

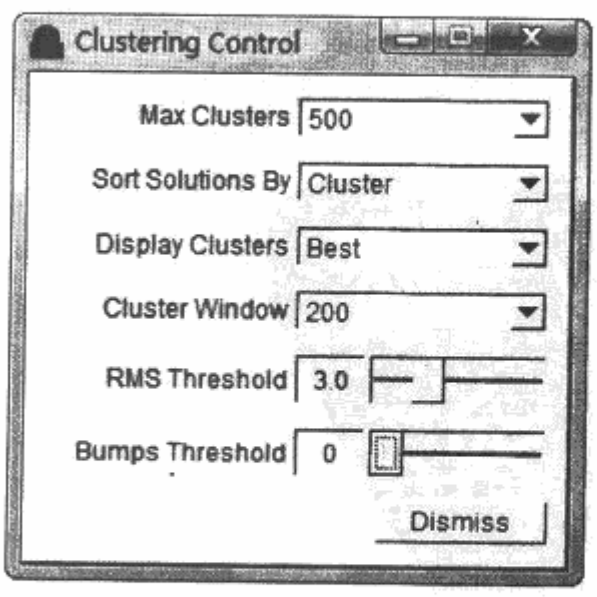


图 6-22　Contorls Clustering

3．保存对接结果

获得 Hex 的对接结果的方法很简单，只需要点击 File | Save 即可。但是，可以看 Save 下有许多的子菜单，这常常让初学者很困惑。实际上，Hex 为用户提供了多种储存对接结

果的方式。以下介绍一下 Hex 储存对接结果的方式。

Receptor：保存当前 3D 窗口中的受体（PDB 格式）。

Ligand：保存当前 3D 窗口中的配体（PDB 格式）。

Both：保存当前 3D 窗口中的配体和受体（PDB 格式）。

Complex：保存 Complex 文件（PDB 格式）。对应于用 File | Open | Complex 读入的 Complex 文件。

Range：保存一定范围内的配体和受体（PDB 格式）。

Summary：保存对接结果的摘要（文本格式），其中包含了能量、聚类等信息。

Matrix：保存配体的变换矩阵（文本格式）。Hex 只能保存而不能读取此类数据。该数据只能用于外部程序，可以通过编写脚本程序来处理该文件，从而获得任一个配体构象的坐标。

Transform：保存能量、配体转动和平移操作的信息（文本格式）。与 Matrix 一样，Hex 不能读取该文件。

三、对接实例

以上介绍了用 Hex 做大分子-大分子对接的流程和方法，现在将通过一个完整的实例来向读者阐述 Hex 在实际科研工作中的应用。为了说明 Hex 的对接效果，选择 Hex 自带的例子来进行说明，操作步骤如下。

① 读入受体、配体及复合物文件（在/Hex 5.1/examples 目录下），文件名分别为 3hfl_fv.pdb、3hfl_ly.pdb 和 3hfl_fv_ly.pdb。

② 设置对接参数，将 Correlation Type 设置为 Shape+Electrostatics，Receptor/Ligand Range 设置为 45°，其余参数均设置为默认值。

③ 检查参数，确认无误后开始对接。

④ 待对接结束后，首先利用 File | Save | Summary 来输出对接结果的基本信息（包含能量、聚类及与参考结构的 RMSD）。

⑤ 综合考虑聚类及打分挑选出最好的结果，并与复合物结构进行比较。

结果：挑选出的最好的对接结果与复合物晶体结构吻合的相当好，其 RMSD 只有 0.74Å（读者计算出的数值可能会略有不同）。

第七章

蛋白质结构的分子动力学模拟

第一节　蛋白质分子动力学模拟概述

蛋白质的生物功能与结构动力学特征紧密相关，不同结构运动模式决定了蛋白质在生化过程中的不同生物功能。蛋白质分子的内部运动和构象动态的变化对于蛋白分子的正常功能过程是非常重要的，这些过程包括蛋白的折叠与伸展、与配体或底物的结合、催化反应、分子识别等过程。在这些不同过程中蛋白质分子的构象变化主要是从精细的局部运动到整体结构构象的变化不等。要了解蛋白质的特性并找到其发挥生物功能的分子基础模型，只有氨基酸顺序和三维结构是不够的，还必须研究蛋白质的结构和功能的关系、蛋白质之间相互作用、蛋白质和具有生物活性的小分子之间的相互作用、蛋白质的构象变化及溶剂对蛋白质构象的影响、无机离子在蛋白质生物过程中的作用等。

目前，测定蛋白质空间结构的实验方法如 X 射线晶体学、多维核磁共振波谱学技术等可以发现当蛋白处于不同的功能状态时对应的不同构象，但对于这些已知不同构象之间的变化路径的细节仍旧是不清楚的。就蛋白质结构动力学研究手段而言，目前还没有成熟的实验技术，因此，利用计算机模拟手段研究蛋白质动态结构与功能的关系就显示出了重要性。随着分子生物学的发展和计算机技术的长足进步，运用各种数学、物理学的理论方法，以计算机为工具进行生物学的基本理论研究，尤其是进行生物大分子体系及其相互作用的计算机分子模拟研究，已经成为十分活跃而且十分重要的研究领域。

计算机对生物分子模拟的结果对于在分子、原子乃至电子结构层次上了解生命现象的基本过程具有极为重要的意义。给定一种生物大分子的空间结构，使用分子动力学模拟计算方法，可以模拟得到分子系统的构象和变化起伏随着时间变化的各种信息，进而得到分子结构的运动轨迹，或者将结构整体运动分解为不同频率振动模态的叠加，能够帮助进一步理解蛋白质生物活性点在配体结合、催化过程中的运动特征。此外，计算机模拟为天然生物大分子的改性和基于受体结构的药物分子的设计提供了重要依据，在某些领域已成为实验研究难以替代的手段。在过去的 10 年中，随着计算机软硬件技术的快速提升，生物大分子的计算机模拟已经成为在原子水平上研究蛋白质、核酸等生物大分子的结构和动力学性质的一个重要工具。这种方法现在已经大量用于研究生物分子以及其复合体的结构、动力学和热力学过程；得到生物分子系统的各种属性，包括蛋白质的稳定性、构象变化信息、蛋白质折叠信息、分子识别（蛋白质、DNA、质膜、复合体）、生物系统内的离子运输等。可以应用于药物设计、X 射线衍射和核磁共振分子结构判别等。

1957 年，Alder 和 Wainwright 首先采用分子动力学（MD）并应用硬球模型来研究凝聚态系统的气体和液体。1977 年，McCammon 等人在“Nature”上发表第一篇蛋白质的分子动力模拟研究论文，研究对象是当时刚刚解析出晶体结构的牛胰岛素抑制剂（BPTI）。

虽然模拟时间很短，但是使人们看到分子动力学技术对研究蛋白质分子功能的重要作用。之后，随着计算机软硬件技术的快速提升，生物大分子的分子动力学模拟已经成为在原子水平上研究蛋白质、核酸等生物大分子的结构和动力学性质的一个重要工具。随着分子模拟技术的飞速发展，逐步形成了一些较完善的软件。应用于生物大分子领域的分子模拟软件主要有美国 Accelrys 公司的 Discovery Studio（即原来的 InsightII）软件包以及 Tripos 公司的 Sybyl 软件；此外，还有一些用于学术研究的模拟软件，如 NAMD、GROMACS 等。以下将介绍 Discovery Studio 软件包中的分子动力学工具以及广泛使用的 NAMD、GROMACS 软件。

第二节　Discovery Studio 平台中分子动力学操作实例

本节将以一个组蛋白-配体复合物将进行带束缚动力学模拟为实例介绍 Discovery Studio 中分子动力学工具的详细操作，Discovery Studio 中的分子动力学采用的是CHARMm 程序，主要过程包括：①准备蛋白-配体复合物；②选择并定义原子组；③添加溶剂体系；④设置和删除束缚部分；⑤进行带束缚的分子动力学模拟；⑥分析动力学模拟轨迹；⑦计算蛋白-配体相互作用能量。

一、准备分子动力学体系，执行分子动力学计算

1．导入蛋白-配体复合物，并赋予力场参数

双击图标，启动 Discovery Studio，在文件浏览器（Files Explorer）找到 1IE9_VDX.dsv 文件，双击打开，将蛋白-配体复合物导入 3D 窗口（图 7-1）。

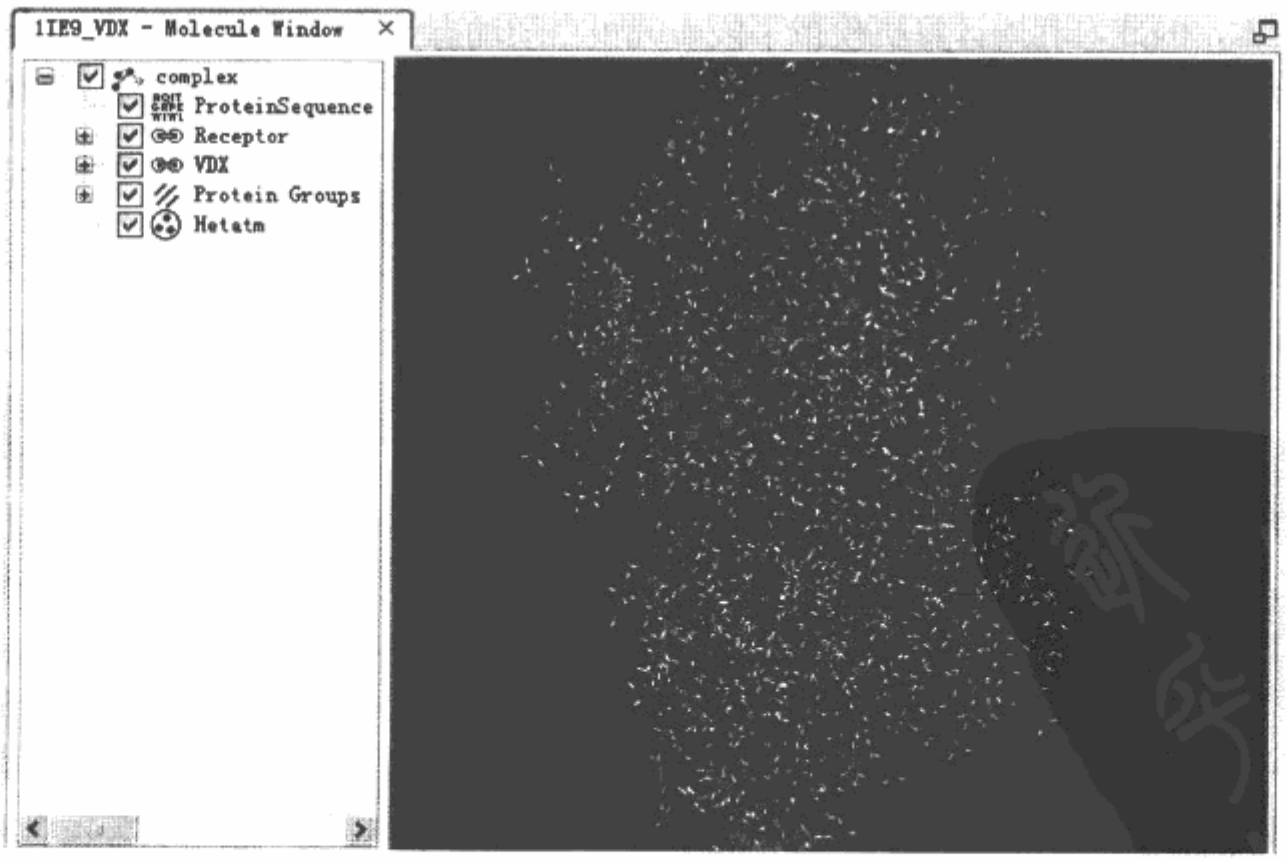


图 7-1　复合物示意图

在工具（Tools）窗口中，鼠标点击窗口最上方列表框的下拉箭头，选择 Simulation，在工具窗口中展示分子动力学模拟工具面板组。

从工具面板组中找到 Simulate Structures | Forcefield 工具面板，确保当前力场为缺省的

CHARMm 力场，否则可通过 Choose forcefield 命令选择 CHARMm 力场，然后点击 Apply Forcefield 命令，给蛋白-配体复合物赋力场（图 7-2）。

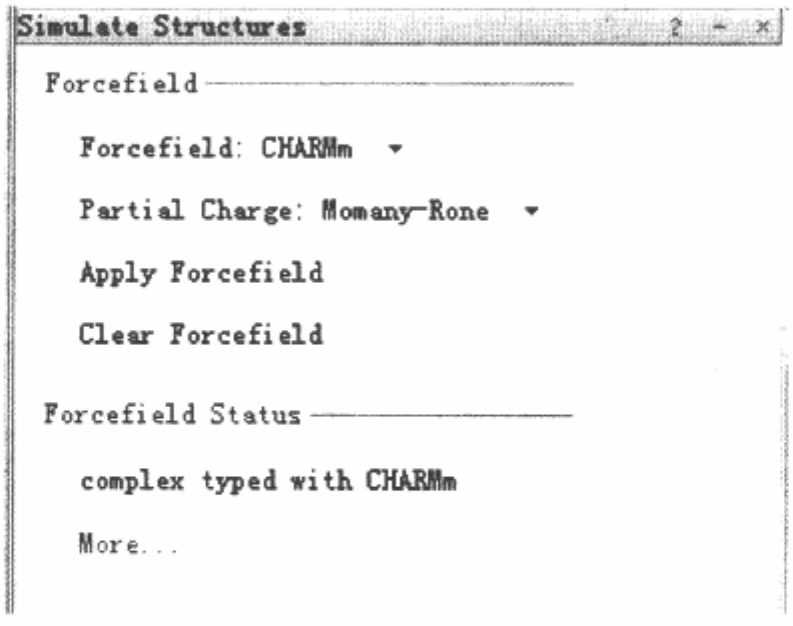


图 7-2　赋与复合物力场参数

2．添加溶剂环境

为使动力学模拟过程更接近真实情况，可模拟人体内环境添加溶剂环境。

流程中，点击“+”号展开 Simulations 文件夹，双击 Solvation 流程，该流程在参数中打开，1IE9_VDX:complex 自动载入到 Input Typed Molecule 参数格。

展开 Solvation Model 选项，设置 Solvation Model 选项为 Explicit Periodic Boundary。这一操作设置添加周期性边界溶剂模型。设置 Cell Shape 选项为 Orthorhombic，此操作设置添加的溶剂盒子为长方体形状。

设置 Add Counterion 为 True，展开 Add Counterion 选项选择 Salt 溶液类型。Salt Concentration 采用缺省值 0.145，阳离子类型（Cation Type）设置为 Sodium，阴离子类型（Anion Type）设置为 Chloride，以上操作表示为复合物添加生理盐水溶剂环境（图 7-3）。

Solvation

Parameter Name	Parameter Value
Input Typed Molecule	1IE9_VDX:complex
Solvation Model	Explicit Periodic Boundary
Radius of Sphere	20.0
Center of Mass	1IE9_VDX:complex
Minimum Distance From Boun…	7.0
Cell Shape	Orthorhombic
Add Counterion	True
Salt Concentration	0.145
Random Seed	314159
Cation Type	Sodium
Anion Type	Chloride

图 7-3　Solvation 参数设置

在工具栏中点击 ▶ 按钮开始添加溶剂，这个作业需要数分钟完成。

下拉菜单中选择 Window | Close All 关闭所有窗口。

当任务完成对话框打开，点击 OK 关闭。双击该作业行打开 Output 文件夹和 Report.htm

文件。点击 complex.dsv 文件，查看添加溶剂环境结果（图 7-4）。

3．选择束缚的氨基酸残基

由于添加溶剂时，水分子和离子都是规则排列，此实例中，在进行动力学时首先束缚住蛋白-配体复合物体系，仅优化一下水分子和离子，所以需要设置束缚体系。

系统视图中点击选中受体分子 Receptor，按住 Shift 键同时选择配体分子 VDX，下拉菜单中选择 Edit | Group…命令，在 Group 对话框中输入 Group 的名字“Fixed”，点击 Defined 定义名字。点击 3D 窗口任意空白处取消选择。新定义的组将出现在系统视图中（图 7-5）。

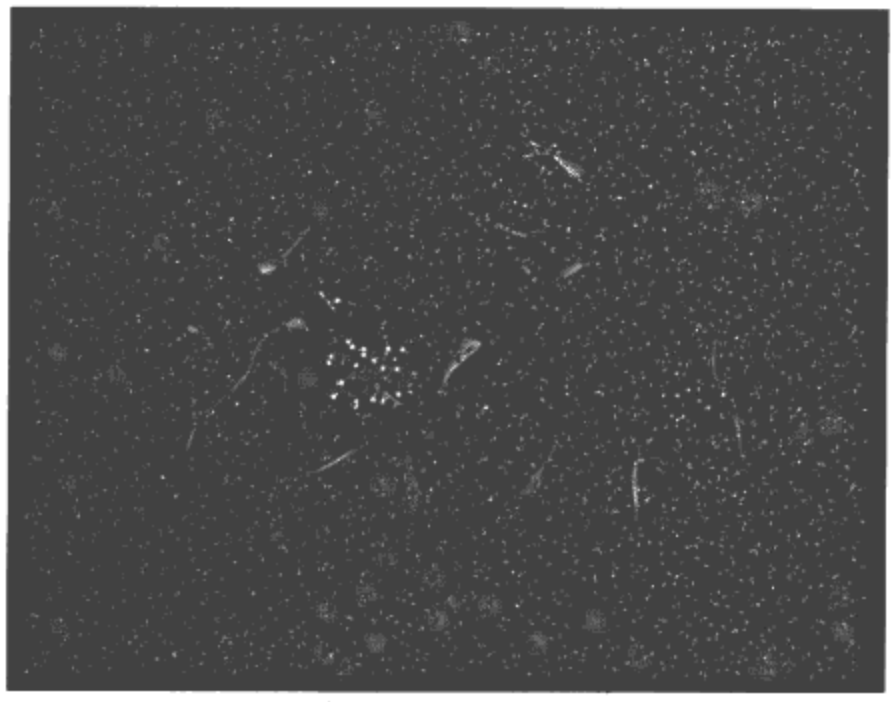


图 7-4　添加溶剂后示意图

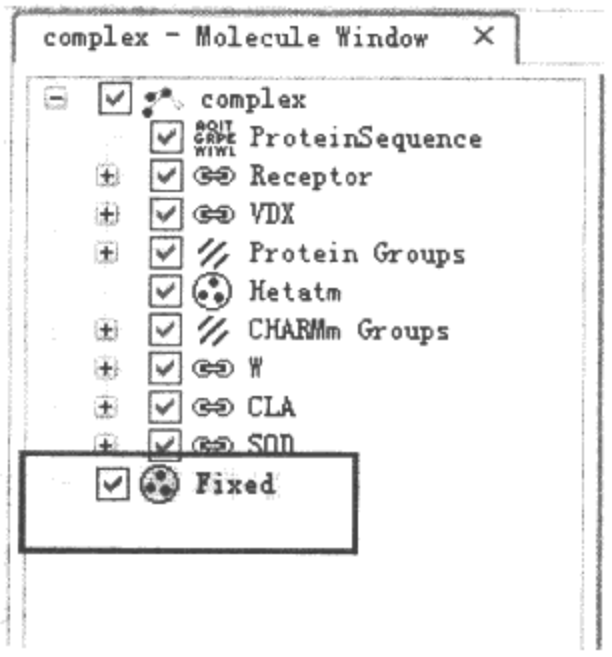


图 7-5　设置束缚体系

4．设置束缚

此实例中，欲先仅仅优化一下水分子及离子，因此，必须束缚体系中其他的原子。

在系统视图中，点击选择 Fixed 组，工具中打开 Simulation | Simulate Structures 下 Constraints 工具面板，点击 Create Fixed Atom Constraint，点击 3D 窗口任意空白处取消选择。3D 窗口中可以看到被束缚的原子周围环绕着蓝色的小槛（图 7-6）。

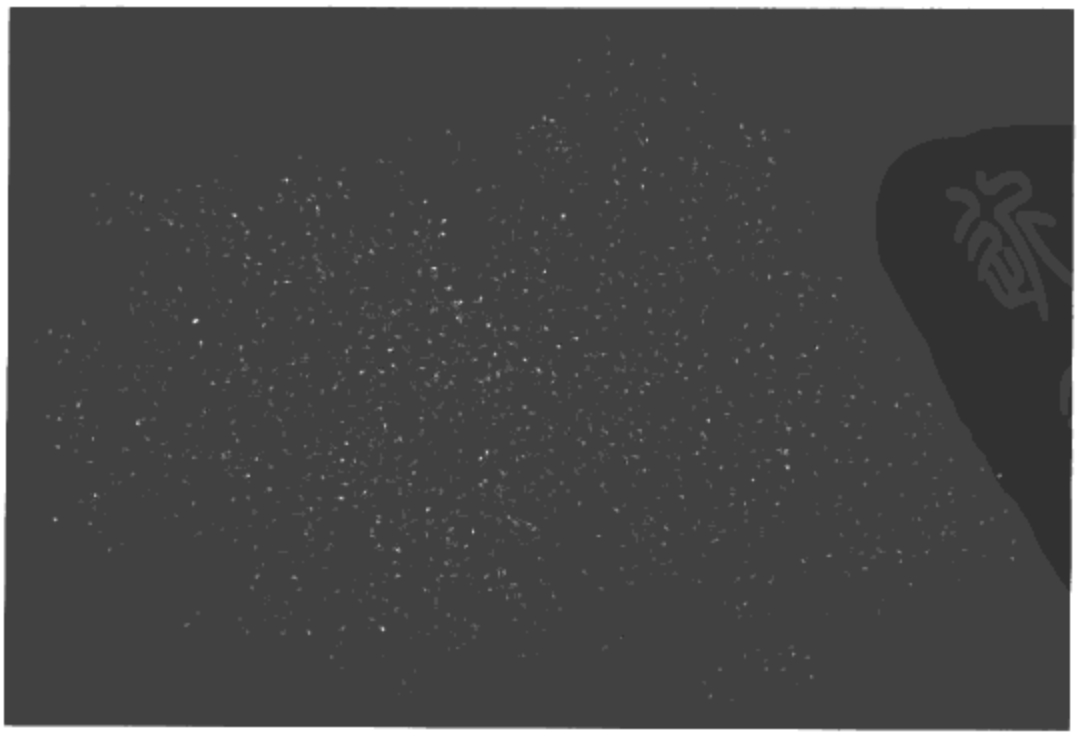


图 7-6　束缚体系示意图

5．打开动力学流程，修改参数

确保添加了溶剂体系的 complex.dsv 窗口为当前活跃窗口，文件中双击 Simulations 文件夹下 Standard Dynamics Cascade 流程，该流程在参数中打开，complex:complex 自动载入到 Input Typed Molecule 参数格。Standard Dynamics Cascade 流程包括 5 个阶段：优化Ⅰ（Minimization）、优化Ⅱ（Minimization2）、加热（Heating）、平衡（Equilibration）、动力学（Production）。展开 Minimization 阶段的参数组，确保 Constraints 选项选择 Fixed:fix_1_Fixed，确保其他阶段中 Constraints 选项为空白。

展开 Prodution 阶段参数组，设置 Steps 为 5000，步长 Time Step 为 0.002，Save Results Frequency 为 100，Type 为 NPT（图 7-7）。

Standard Dynamics Cascade ×

Parameter Name	Parameter Value
Input Typed Molecule	complex:complex
⊟ Minimization	
Algorithm	Steepest Descent
Max Steps	500
RMS Gradient	0.1
Energy Change	0.0
Constraints	Fixed:fix_1_Fixed
⊞ Minimization2	
⊞ Heating	
⊞ Equilibration	
⊟ Production	
Steps	5000
Time Step	0.002
Target Temperature	300.0
Save Results Frequency	100
Save Restart File	False
Constraints	
Type	NPT
TMass	1000.0
⊞ Implicit Solvent Model	None
⊞ Nonbond List Radius	14.0
⊞ Electrostatics	Particle Mesh Ewald
⊟ Advanced	
Dynamics Integrator	Leapfrog Verlet
Apply SHAKE Constraint	True
Random Number Seed	314159
Number of Processors	1

图 7-7　Standard Dynamics Cascade 流程参数设置

设置 Implicit Solvent Modeler 选项为 None，Electrostatics 选项选择 Particle Mesh Ewald。

在工具栏中点击▶按钮开始分子动力学计算。这个作业需要数分钟完成，作业的状态可以在 Jobs 中查看。

当任务完成对话框打开，点击 OK 关闭。

6．查看结果，显示构象变化动画

双击该作业行打开 Output 文件夹和 Report.htm 文件。点击打开 complex_nowater.dsv 文件，按 Ctrl+1 最大化观测区域，在系统视图中选择 Fixed:fix_1_Fixed，单击鼠标右键，在右键菜单中选择 Hide，隐藏每个原子周围的蓝色小槛。

确保 3D 窗口是当前活跃窗口，选择下拉菜单 Structure | Animation | Play 命令，则可以看到在 3D 窗口中的复合物开始运动，显示动力学模拟过程中的变化，动画循环往复播放。

再次选择下拉菜单 Structure | Animation | Play 命令可以停止动画播放。

二、分析分子动力学计算结果

1．绘制能量曲线

在工具窗口中，鼠标点击窗口最上方列表框的下拉箭头，选择 Simulations，在工具浏览器窗口中展示 Simulations 工具面板组。打开 Analyze Trajectory 工具面板组，选择 Create Plot…命令，打开 Choose X and Y Axes 对话框，设置 Graph Type 为 Line，左侧的 Choose X Axis 栏里选择 Step，右侧的 Choose Y Axis/Axes 栏里选择 Potential Energy，点击 OK 按钮。将会在新窗口中打开能量-时间关系曲线。该曲线只显示了动力学阶段（Production Phase）的能量变化。

在曲线图示区域点击鼠标右键，选择右键菜单中 Display Style…命令，点击 Advanced 键，在面板的底部选择 Show points on line，点击 OK（图 7-8）。

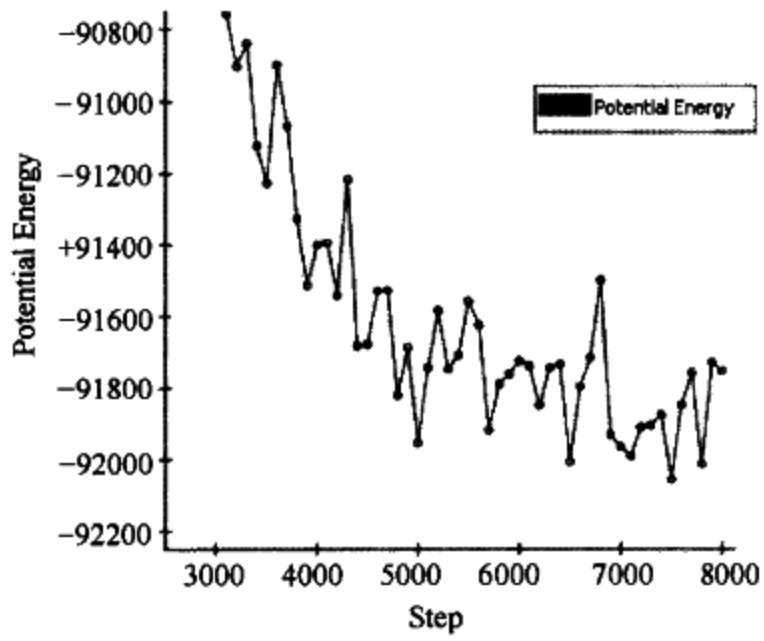


图 7-8　能量曲线示意图

移动鼠标将指针指到圆点上，可以显示对应的步数及电势能。

2．计算相互作用能

采用动力学计算所得的轨迹，可以计算复合物体系中任意一组原子间的相互作用能。此实例中，以计算蛋白-配体之间的相互作用能为例。

确保 Complex_nowater.dsv 窗口为当前活跃窗口，在流程中展开 Simulation 文件夹，双击 Calculate Interaction Energy 流程，该流程在参数中打开，复合物名称 complex:complex 自动载入到 Input Typed Molecule 参数格，另外两个参数选项显示为红色，需要自己定义相应的组。

在系统视图中选择 VDX，然后在参数中点击 Atom Selection 1 参数，点击下拉列表中 Create New Group from selection…条目，则在参数栏中出现 Group 项目。此操作表示选择配体作为第 1 个组。

在系统视图中选择 Receptor，在参数中点击 Atom Selection 2 参数，点击下拉列表中 Create New Group from selection…条目，则在参数栏中出现 Group1 项目。此操作表示选择整个蛋白作为第 2 个组（图 7-9）。

保持 Calculate Interaction Energy 流程中其他参数为缺省设置，在工具栏中点击 ▶ 按钮开始计算，此作业将会计算动力学轨迹中所有构象中的蛋白-配体间的相互作用能。

3．打开相互作用能计算结果

当任务完成对话框打开，点击 OK 关闭。双击该作业行打开 Output 文件夹和 Report.htm 文件。点击 complex.dsv 文件，Ctrl+T 打开表格浏览器（Data Table），选择 Conformation 键显示每个构象单独的数据，拖动表格找到四个新增栏：ID、INTERACTION ENERGY、

ELECTROSTATIC INTERACTION ENERGY 及 VDW INTERACTION ENERGY。

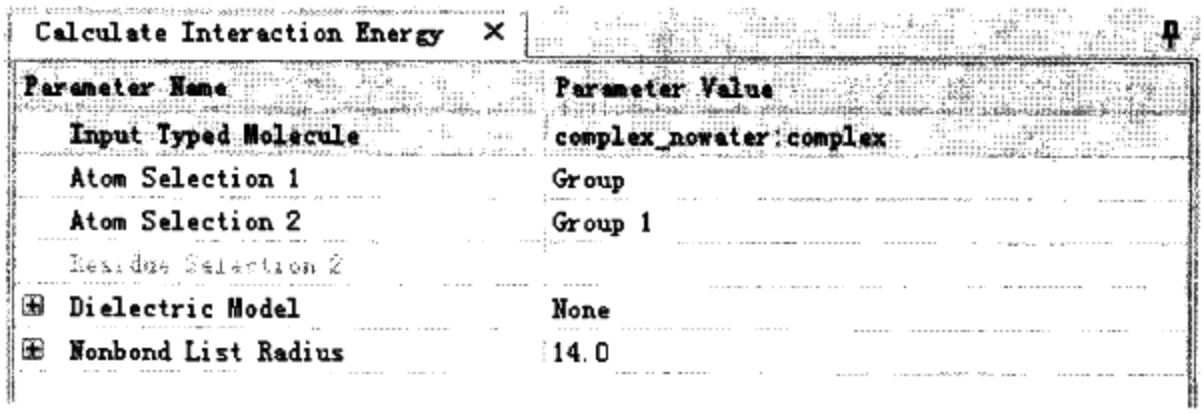


图 7-9　Calculate Interaction Energy 流程参数设置

以上栏目中显示了所选定的两个原子组（在此实例中为蛋白-配体）间的非键作用能量及其各组分的数值。

4．绘制相互作用能曲线

采用表格中的数据，可以绘制数据曲线。

点击选择下拉菜单 Chart | Line Plot 命令，打开 Choose Plot Axes 对话框，对话框左侧的 Choose X Axis 中选择 Step，对话框右侧的 Choose Y Axis/Axes 中选择 INTERACTION ENERGY，按住 Ctrl 键连续选择 ELECTROSTATIC INTERACTION ENERGY 及 VDW INTERACTION ENERGY，点击 OK（图 7-10）。

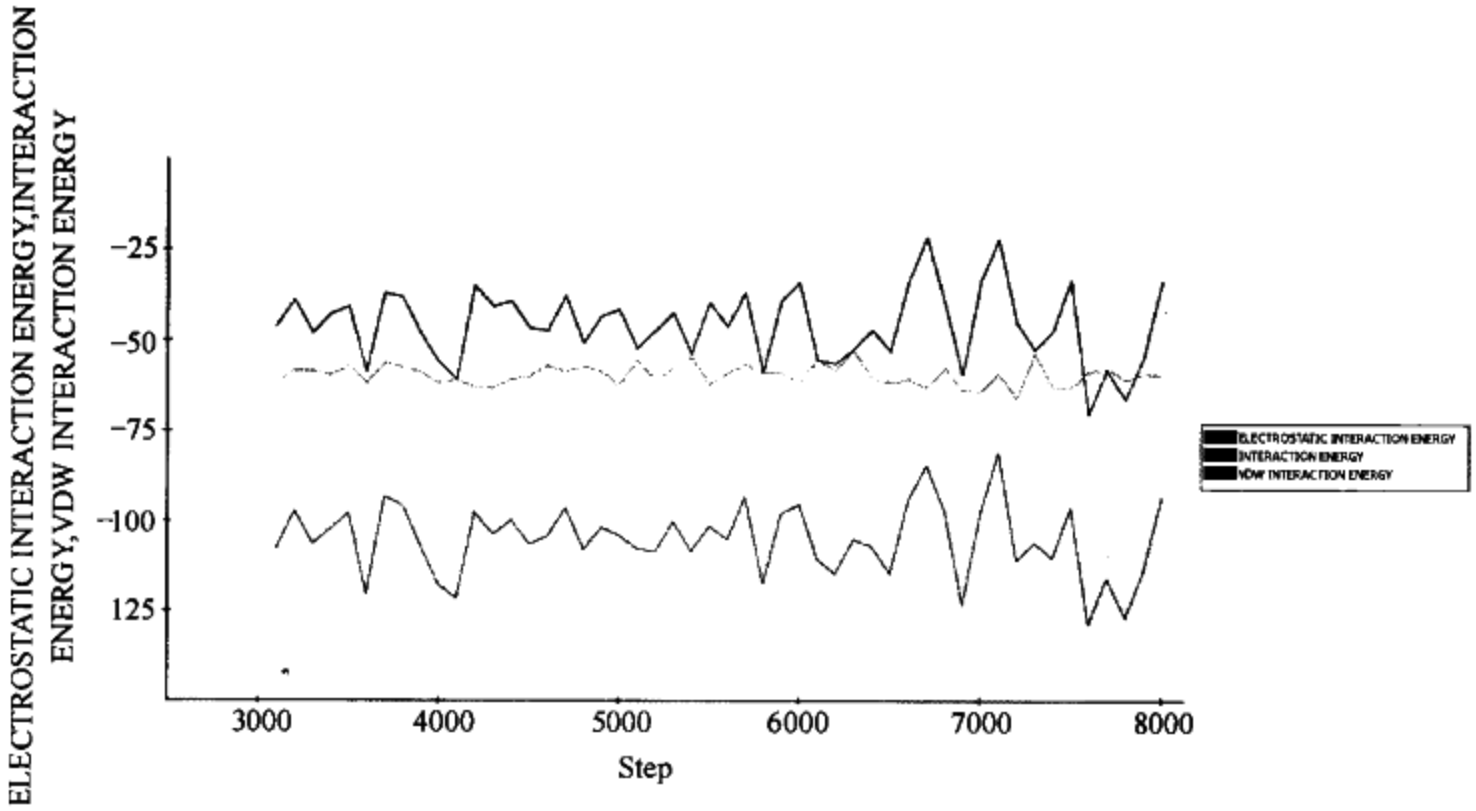


图 7-10　相互作用能曲线示意图

三条能量曲线图将在新窗口中打开，这是检验相互作用能的重要性质。较低的相互作用能预示了合理的结合构象。从能量曲线图中，也可以清楚地看到每一步或每一个构象的范德华和静电能量的分布。

5．保存最低能量构象

可以选择保存动力学轨迹中任意构象以备进一步分析使用。

在 complex – Line Plot 窗口中选择最低能量构象，激活 3D 窗口，打开系统视图，将选中的高亮的构象显示出来，点击选择下拉菜单 Edit | Copy 命令，点击选择下拉菜单 File | New | 3D Window 命令，打开一新的 3D 窗口，在新窗口中选择下拉菜单 Edit | Paste 命令。则选中的构象出现在新的 3D 窗口中，可以用于下一步研究和分析。

6．定义二面角监控

激活 3D 窗口，打开系统视图查看轨迹文件，按住 Ctrl 键选择配体 VDX 中的四个原子：C23、C24、C25、C27，点击选择下拉菜单 Structure | Monitor | Torsion 命令。

绿色的二面角监控将呈现在图形界面里，同时可以看到系统视图（Hierarchy View）及表格下的 TorsionMonitor 里也会出现监控信息。当播放动画时，二面角也会随构象变化而改变（图 7-11）。

7．定义距离监控

在系统视图（Hierarchy View）中选择配体 VDX 中的 O3 以及蛋白受体的 His397 氨基酸残基的 NE2 原子，点击选择下拉菜单 Structure | Monitor | Distance 命令。绿色的距离监控将呈现在图形界面里的两个原子之间（图 7-12）。

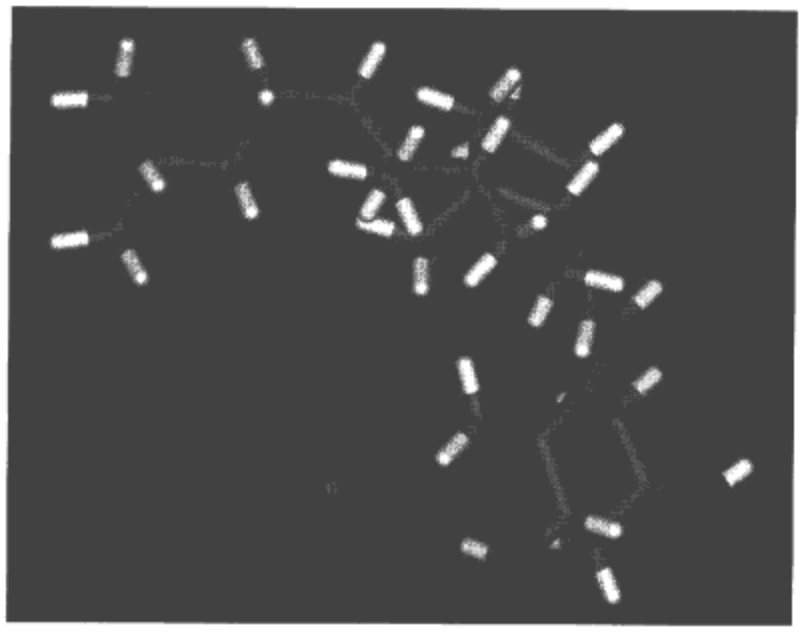


图 7-11　二面角监控示意图

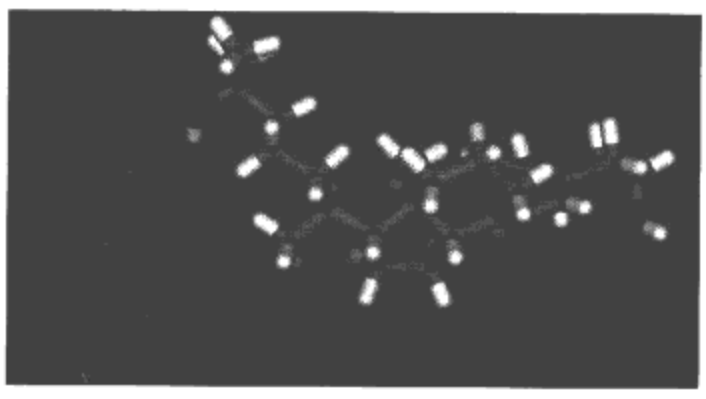


图 7-12　距离监控示意图

8．开启氢键监控

点击选择下拉菜单 Structure | Monitor | HBond 命令。结构中的所有氢键都将以绿色虚线的形式显示出来，当播放动画时，二面角也会随构象变化而改变（图 7-13）。

另外，Intermolecular HBonds 命令将仅显示独立的分子间的氢键。由于此实例中仅包含一个复合物分子，故在此选择 HBond 命令。

9．建立 RMSD 对比关系，编译动力学轨迹监控数据

在流程中展开 Simulation，双击 **Analyze Trajectory** 流程，该流程在参数中打开，复合物名称 complex:complex 自动载入到 Input Molecule 参数格。

点击 Properties 参数，勾选 Torsion1、Distance1 以及 HBond Monitor1。

点击 RMSD Reference 参数，下拉列表中选择 **1** 作为参考。

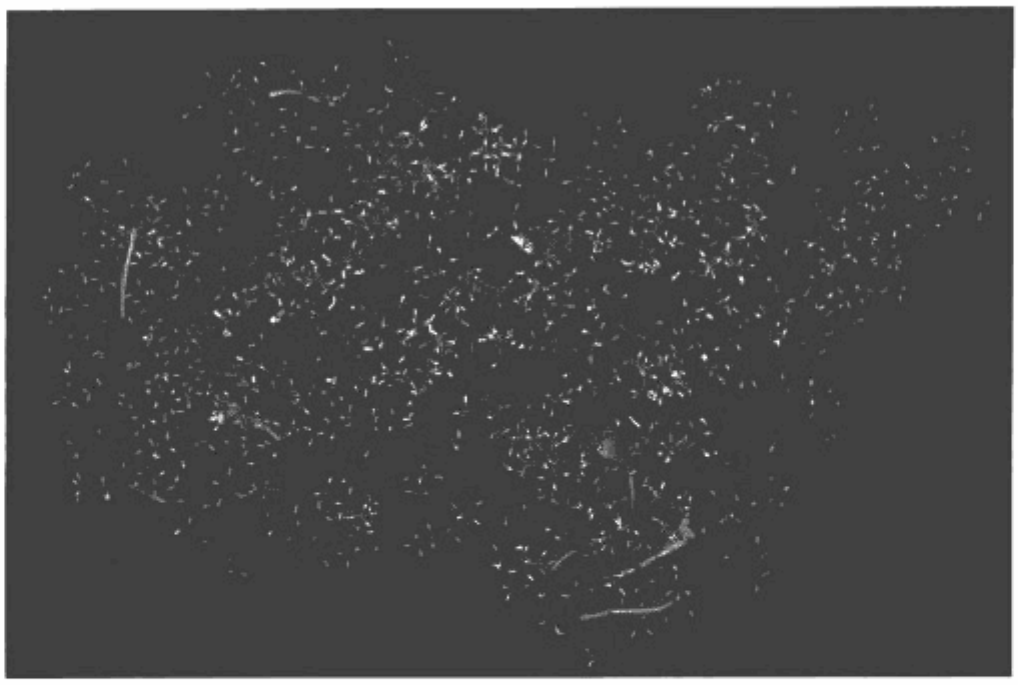


图 7-13　氢键监控示意图

展开 RMSD Reference 参数组，确保 Scope 参数设置为 Molecule，此操作确定计算的是每个分子的 RMSD 数值。

设置 Atom Group 及 Atom Group to Fit 参数为 Backbone（图 7-14）。

Analyze Trajectory ×

Parameter Name	Parameter Value
Input Molecule	complex:complex
Properties	Torsion1, Distance1, HBond Monitor1
⊞ Torsion Angle Properties	
⊟ RMSD Reference	1
Scope	Molecule
Label	RMSD to conf
Atom Group	Backbone
Atom Group to Fit	Backbone
Save Fitted Trajectory	True
⊞ Advanced	

图 7-14　Analyze Trajectory 流程参数设置

在工具栏中点击 ▶ 按钮开始计算。点击选择下拉菜单 Window | Close All 命令，如果询问是否保存文件，选择 No。当任务完成对话框打开，点击 OK 关闭。双击该作业行打开 Output 文件夹和 Report.htm 文件。点击打开 complex.dsv 文件。

10．绘制监控数据曲线

绘图之前，确定表格中包含所需数据。

激活 Complex 3D 窗口，在工具浏览器中选择 Simulation 工具面板，打开 Analyze Trajectory 工具面板，点击 Create Plot…命令。在 Create Plot…对话框中，确定设置 Plot type 为 Line，设置左侧的 Choose X axis 栏为 Time，右侧的 Choose Y axis/axes 中选择 Torsion1。点击 OK 关闭对话框。二面角曲线将在新窗口中生成（图 7-15）。

Analyze Trajectory 工具面板中，点击 Create Plot…命令。在 Create Plot…对话框中，确定设置 Plot type 为 Line，设置左侧的 Choose X axis 栏为 Time，右侧的 Choose Y axis/axes 中选择 Distance1。点击 OK 关闭对话框。距离曲线将在新窗口中生成（图 7-16）。

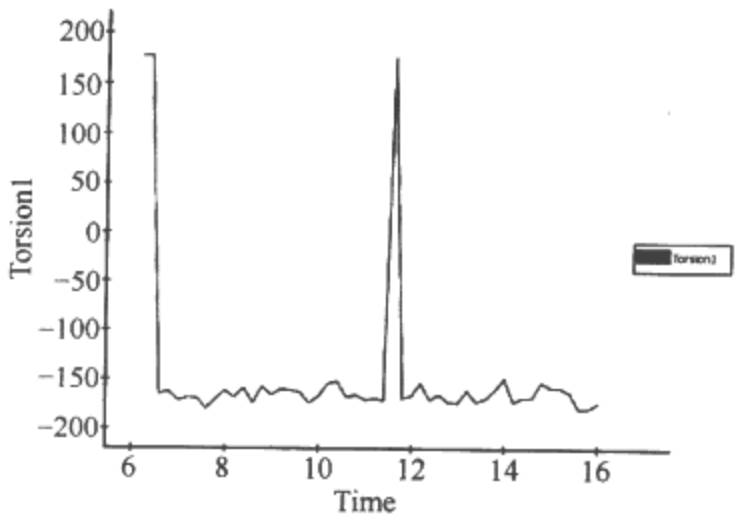


图 7-15　二面角监控曲线示意图

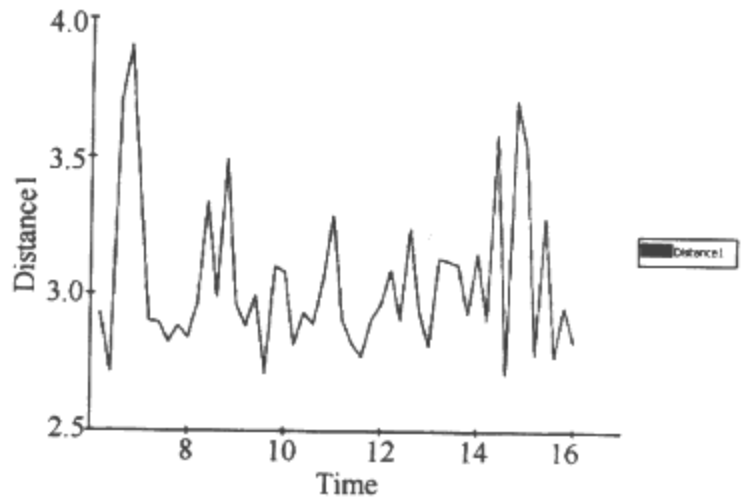


图 7-16　距离监控曲线示意图

11．绘制氢键热图

激活 Complex 3D 窗口，打开表格，选择 Conformation 标签，滚动表格至氢键监控数据，HBond Monitor1 纵栏中显示了该构象中找到的氢键总数。

选择标有纵栏，按住 Shift 键选择氢键最后一纵栏，此操作选择了除 HBond Monitor1 外，所有氢键相关栏目。

点击选择下拉菜单 Chart | Heat Map 命令，氢键热图将在新窗口中打开，此图显示了动力学过程中复合物体系中的氢键情况（图 7-17）。

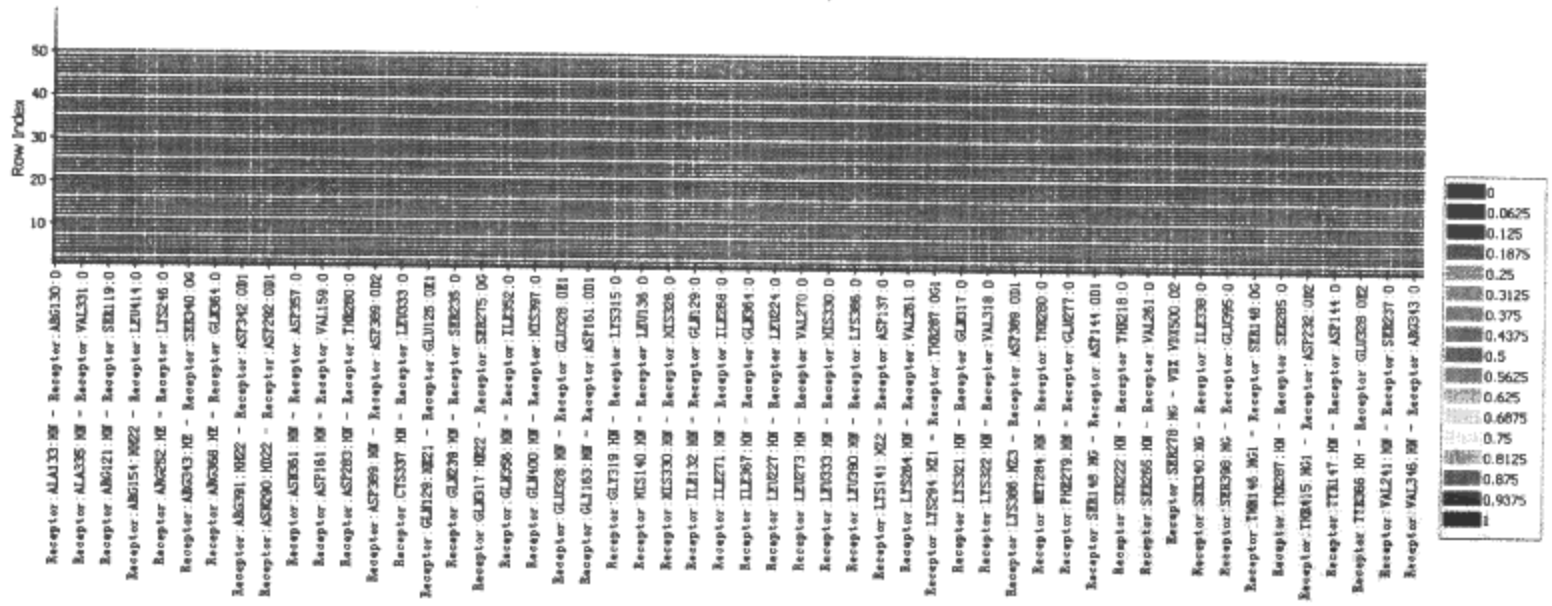


图 7-17　氢键热图示意图

12．绘制 RMSD 曲线

激活 Complex 3D 窗口，在工具浏览器中选择 Simulation 工具面板，打开 Analyze Trajectory 工具面板，点击 Create Plot…命令。

在 Create Plot…对话框中，确定设置 Plot type 为 Line，设置左侧的 Choose X axis 栏为 Time，右侧的 Choose Y axis/axes 中选择|RMSD to conf 1。点击 OK 关闭对话框。RMSD 曲线将在新窗口中生成 （图 7-18）。

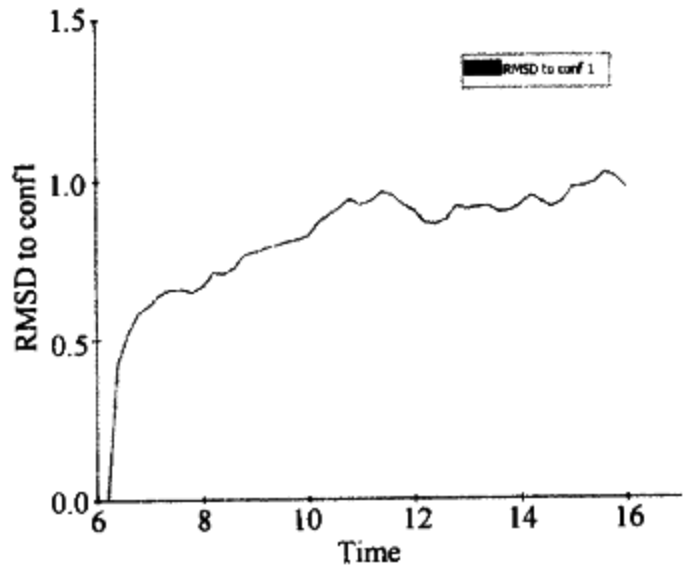


图 7-18　RMSD 曲线示意图

第三节　利用 NAMD 进行分子动力学计算的操作及研究实例

一、NAMD 简介

NAMD 是一款由美国伊利诺伊大学厄本纳-香槟分校（University of Illinois at Urbana-Champaign，UIUC ）贝克曼研究所理论生物物理小组开发的优秀的分子动力学软件。它计算速度快，运行效率高，能在多种软硬件平台下运行，同多种分子动力学软件和力场兼容。且使用简便，用户群大，广泛用于科研和生产之中。NAMD 本身只能完成动力学计算功能，建模和数据处理要在其他软件中进行，这些辅助软件中最重要的是 VMD。VMD 是可以完成一款视图、建模、数据处理等多种功能。它同 NAMD 系出同门，已实现同 NAMD 的无缝链接。NAMD 和 VMD 还提供了一个 TCL （tool command language）语言的接口，可以通过 TCL 编程实现若干功能而不需要阅读和修改程序的源代码。NAMD 对 CHARMm 力场兼容性最好，而这个力场主要是针对核酸、蛋白质和脂类。同时 VMD 的大部分功能也主要是针对生物体系。鉴于这些方面的限制，NAMD 目前多被用来模拟生物体系。对于非生物体系的研究如碳纳米管、SiO_2 等也有报道。NAMD 和 VMD 的下载地址分别为：http://www.ks.uiuc.edu/Research/namd/ 和http://www.ks.uiuc.edu/Research/vmd/。

二、NAMD 程序的安装与使用

VMD 和 NAMD 同时提供源代码版本和多种平台下的预编译版本，用户可以根据需要选择安装。预编译版安装方便，且使用了大量的优化选项进行了编译，运算速度较快，如果没有特殊要求可以安装这种。当机器平台不在预编译版本的支持之列，或是通过删改源代码的方式增加或改进程序的功能时，就必须自己编译安装了。

1．安装方法

以下介绍 VMD 和 NAMD 的预编译版本的安装过程。所有演示都是在 RedHat® Enterprise Linux 5 和 Microsoft® Windows® xp 下完成，Linux 的 shell 为 bash 3.1。

LINUX 下，解压 VMD 安装文件后进入，输入/configure　Linux，再进入 src 目录，输入 make　install，如果出现”VMD installation complete.　Enjoy!”字样，说明 VMD 安装完成。程序文件默认安装到/usr/local/bin 目录，所以这个操作需要 root 权限。修改

configure 文件更改安装路径后，可以以普通用户权限安装。NAMD 为 tar.gz 压缩文件，解压后无须安装即可使用。

Windows 下，双击 VMD 安装文件即可完成安装，默认安装路径为 C:\Program Files\University of Illinois\VMD，启动快捷方式可以在开始菜单中找到。NAMD 为 zip 压缩文件，解压后无须安装即可使用，比如可以解压后放到 C:\Program Files\University of Illinois\NAMD 目录中。

安装包内有如下几个重要文件:

charmd, charmrun: charm++ 的守护和调用程序,用以在并行环境中使用 NAMD。

namd2:为 namd 的主程序，除完成动力学计算的功能外，还负责进行消息传递和任务分发等。

psfgen: psf 文件的生成程序，用于体系建模。

flipbinpdb, flipdcd: 二进制轨迹文件在 big-endian 和 small-endian 机器之间转换。

还有一些很有用的文件没有随软件包一起发布，如 catdcd 等,它可以很方便编辑轨迹文件。这些可以到http://www.ks.uiuc.edu/Research/vmd/plugins/下载。

VMD 在某些 Linux 发行版下因为系统库依赖关系无法正常使用，最常见的问题是缺少 libstdc++.so.5，表现为程序启动后立即退出，在终端没有留下报错信息。安装下面这个库文件可以解决这一问题。

debian/ubuntu : apt-get install libstdc++5

redhat/suse/redflag : 安装 compat-libstdc++-5.0.7-xxx.rpm

或是在其他机器上寻找一个 libstdc++.so.5 库文件后放入/usr/lib

设置环境变量:

在 Linux 下，在 bashrc 文件中加一行 PATH=$PATH:/namd 存放目录/;

在 Windows 下,右击我的电脑>属性>高级>环境变量>在系统变量中找的 path>在后面加分号再把目录加上>确定。

源代码编译过程较为复杂，可以参考随代码一起发行的 notes.txt 文件。

2．使用方式

NAMD 在一个计算核心上的使用方法很简单，只需要准备好一个配置文件，然后运行 namd2 <configure file> 即可。

VMD 是一个分子动力学视图和数据分析的软件，有多种使用方式。

（1）图形界面 在控制台下输入 VMD，可以以图形界面方式启动 VMD。启动后包含 3 个窗口 （见图 7-19）：main 窗口（左上）、console 窗口（左中）和 display 窗口（右）。main 窗口中有大量的菜单和按钮，用于进行基于图形界面的常用操作；console 窗口用于 VMD 的命令行操作，也用于输出信息；display 窗口用于显示分子结构。

（2）命令行交互模式 VMD 内嵌有 tcl，在 tcl 命令行模式下 VMD 可以实现大量而灵活的功能。有些功能不在 main 窗口的菜单中，只能在这种模式下进行。命令行操作可以在 console 窗口中进行，也可以打开 Extensions | TkConsole，启动 tk 终端（左下）。后者有自动补全、语法高亮等功能，使用起来很方便。如果用 vmd -dispdev text 则可以以文本方式启动 VMD，只会出现 console 窗口，可以用于远程登录时候使用。

（3）脚本模式 把在交互模式下的命令按顺序写入一个文件中，VMD 可以以脚本方

式运行此文件。使用流程控制语句可以实现更强大的功能。启动方式为 vmd –dispdev text –e *<script file>* 或 vmd –dispdev text < *<script file>*。

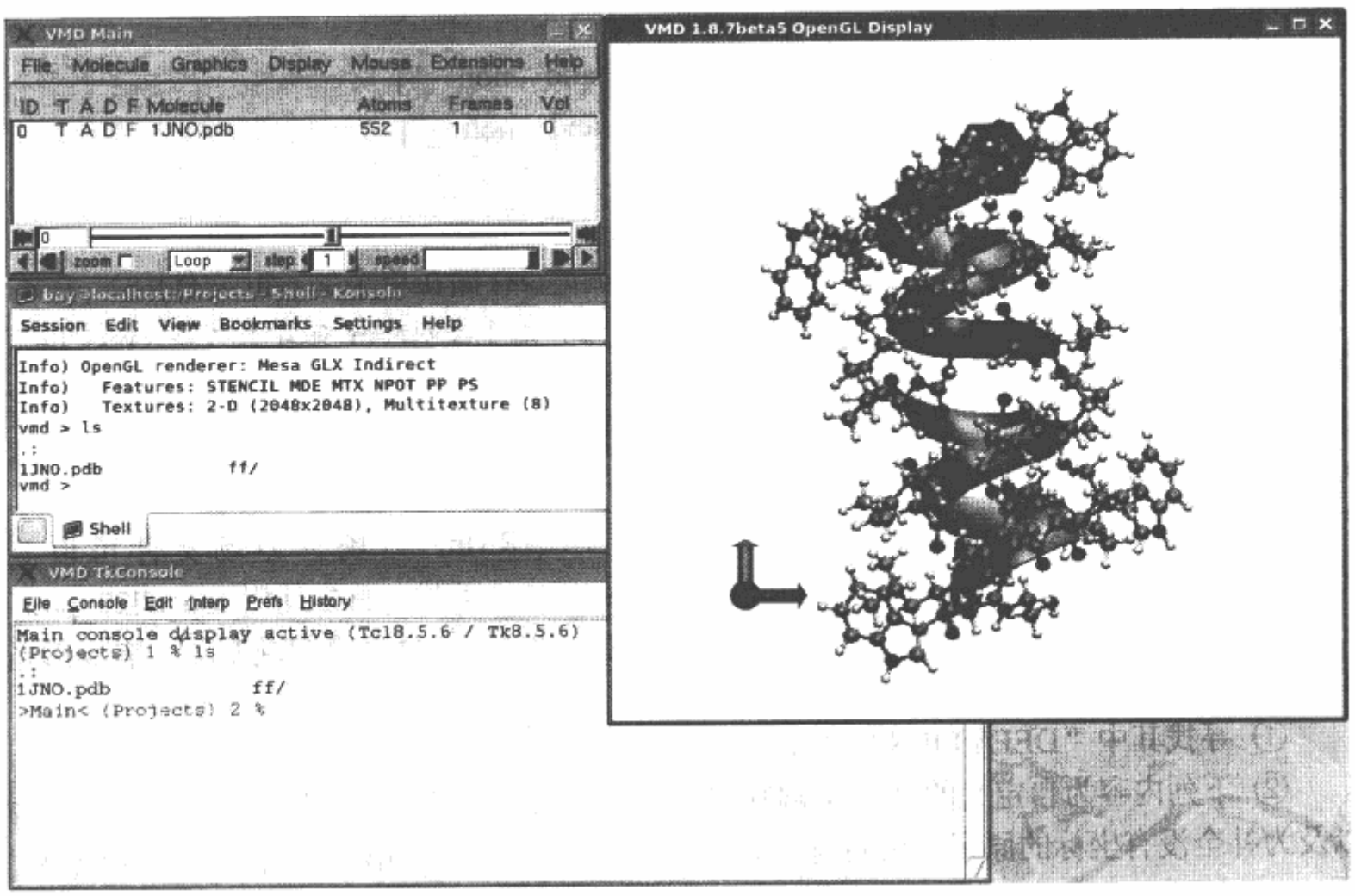


图 7-19　VMD 程序界面

三、构建模型

以短杆菌肽 A（gramicidin A，GA）为例，介绍使用 NAMD 进行分子动力学模拟的步骤。GA 是由 2 个包含 15 个氨基酸残基的肽链形成的管状二聚体，在细胞膜中作为传输水和 Na^+、K^+的通道。GA 单聚体序列为 VGALAVVVWLWLWLW，其中带下划线的表示 D-氨基酸残基，N 端由甲酰基（FOR）封端，C 端由乙醇氨基（ETA）封端。形成中空管状结构，管长 24Å，内径 4～5Å，分子结构如图 7-19 所示。因其结构和功能简单，在确定其结构之后就作为分子动力学模拟的热门分子。

但当进行分子建模时，因为其中包含多条肽链、D-氨基酸残基和不常见的链端封端基团，并且要为该蛋白质建立一个膜环境，这个例子涉及的问题是比较多的。

1．获取 GA 的结构文件

这可以从 pdb 蛋白质结构数据库（http://www.rcsb.org）下载，pdb 号为 1JNO。

2．查看分子结构

以图形界面方式启动 VMD，File | New Molecule | Browse，选择 pdb 文件 | Load，载入分子（在 VMD 中，称存储在一个文件中的体系为一个分子），可以在 display 窗口中查看分子结构。打开菜单 Graphics | Representation | Drawing style | Drawing Methods，可以选择多种视图模式查看。VMD 可以同时载入多个分子（文件），在 main 窗口中间的框中为载入分子的列表。每行最前面一列为文件的编号。这个编号从 0 开始，每个分子唯一，作为它的标识。

3．分割文件

打开 TkConsole 窗口，输入如下命令把 GA 分子的每条链分别存为 1 个文件，并略去其中所有 H 原子。

```
set chainA [atomselect 0 " chain A and noh "]
set chainB [atomselect 0 " chain B and noh "]
$chainA writepdb chainA.pdb
$chainB writepdb chainB.pdb
```

在用 psfgen 生成 psf 文件的时候，每个 segment 要存储在单个的 pdb 文件中。

在 NAMD 中，每条肽链作为一个 segment 单位，每个 segment 有若干残基（residue），每个残基有若干原子。这是 NAMD 划分的三层结构。体系内全部的水和脂可分别作为一个 segment，或是把脂的上下两层作为两个 segment，每个水分子或脂分子作为一个 residue。

4．建立适合 GA 的力场文件

NAMD 兼容多种力场，使用最多的是 CHARMm 力场。力场文件可以到 http://mackerell.umaryland.edu/CHARMm_ff_params.html 处下载。首先使用 top_all27_prot_lipid.inp 构建 GA 分子的 psf 文件。但由于这个文件中并不包含 2 个封端基团，需要适当修改。

① 寻找其中“DEFA FIRS NTER LAST CTER”一行，注释掉。

② 下列内容为自定义的 2 个封端基团的结构信息，包括原子在力场中的类型、电荷以及为补全没有坐标的原子而加的内坐标信息。这些内容在文件最末 END 标记之前加入。

```
RESI FOR          0.00
GROUP
ATOM HA           HA        .0900  !
ATOM C            C         .4200  !
ATOM O            O        -.5100  !
BOND C HA C +N
DOUBLE O C
IMPR C HA +N O
IMPR C +N HA O

RESI ETA          0.00      ! ethanolamine in peptide chain
GROUP                       !
GROUP
ATOM N     NH1    -0.47     !          |
ATOM HN    H       0.31     !          N-HN
ATOM CA    CT2    -0.02     !          |
ATOM HA1   HA      0.09     !          |
ATOM HA2   HA      0.09     !     HA1-CA-HA2
GROUP                       !          |
ATOM CB    CT2     0.05     !          |
ATOM HB1   HA      0.09     !    HB2---CB---HB1
```

```
ATOM HB2   HA      0.09   !          |
ATOM OG    OH1    -0.66   !          OG---HG1
ATOM HG1   H       0.43   !
BOND N HN N CA
BOND CA HA1 CA HA2 CA CB
BOND CB HB1 CB HB2 CB OG
BOND OG HG1
IMPR N -C CA HN
IC-C    CA   *N    HN     1.3474  124.37  180.00  114.18  0.9999
IC N    CA   CB    OG     1.5084  107.86   48.44  104.92  1.3987
IC CA   CB   OG    HG1    1.5191  104.92  173.19  111.99  0.9495
IC HN   N    CA    CB     1.0145  108.35  -49.03  107.86  1.5191
IC N    CB   *CA   HA1    0.0000  000.00  120.00  000.00  0.0000
IC N    CB   *CA   HA2    0.0000  000.00 -120.00  000.00  0.0000
IC OG   CA   *CB   HB1    0.0000  000.00  120.00  000.00  0.0000
IC OG   CA   *CB   HB2    0.0000  000.00 -120.00  000.00  0.0000
```

top 文件是 CHARMm 力场的内容之一，在其中定义了氨基酸残基和脂分子中各个原子的类型、电荷、连接方式等。用以在 psfgen 中调用，生成体系的 psfgen 文件。关于 top 文件的格式可以参考 namd-tutorial（http://www.ks.uiuc.edu/Training/Tutorials/）的附录三。

5．制作 psf 文件

首先把 chainA/B.pdb 中包含原子坐标的最后 2 行移至最前。然后在 TkConsole 中依次输入，如下指令。

```
package   require   psfgen
topology   top_all27_prot_lipid.inp
pdbalias   residue   DLE   LEU
pdbalias   residue   DVA   VAL
pdbalias   atom   ETA   O   OG
segment   GA1   {   pdb   chainA.pdb   }
segment   GA2   {   pdb   chainB.pdb   }
coordpdb   chainA.pdb   GA1
coordpdb   chainB.pdb   GA2
guesscoord
writepdb   protein.pdb
writepsf   protein.psf
```

运行结束后在当前目录下生成 protein.pdb/psf 文件。

在 VMD 和 psfgen 中同时有一个命令叫 writepdb，且功能相同。差别在于 VMD 中的在前面以一个变量开始，在 psfgen 中直接开始。writepdb 可以覆盖原来的 pdb 文件而不给出任何提示，如果对操作的正确性没有把握。建议保存为新的文件名。当确认无误后改回。

psf（protein structure file）文件主要内容为定义了体系中各个原子的连接方式，主要有键、角、二面角等，即可用于 VMD 中显示成键与否，又是 NAMD 在进行动力学计算

的必要信息。在 NAMD 运行时候，它存储的是动力学过程中的不变量，同时从 pdb 文件读取坐标。

psfgen 有 2 种调用方式:

① 随 NAMD 安装包一起发行的有 psfgen 文件，此文件在控制台终端使用；

② VMD 中有一个 plugin 名叫 psfgen，在 TkConsole 中调用。调用前需要运行 package require psfgen。

因后者具有自动补全、语法高亮等功能，且能与 VMD 中的其他功能同时使用，推荐这种方式。

guesscoord 为自动补全缺失原子的命令。蛋白质结构获得的主要手段是 X 射线衍射（XRD）或 NMR，这两种技术都不能精确定位 H 原子，所以在原始的 pdb 文件中一般并无氢原子的信息。psfgen 根据 top 文件中定义的内坐标信息（IC 开头的行）和简单猜测来补全这些缺失原子的坐标。

6. 构建膜

VMD 中有一个 plugin 名叫 membrane，可以快速生成一个双分子层质膜结构。但目前只支持 POPE 和 POPC 两种质膜模型。在 VMD 中的 TkConsole 中输入：

```
package require membrane
membrane -l POPE -x 50 -y 50
```

此命令可以生产 50 Å × 50 Å 的 POPE 膜，在当前目录下生成名为 membrane.pdb/psf 的文件。

7. 调整结构

由 membrane plugin 生成的膜的 x、y 坐标中心不在原点，需通过移动使之到达坐标原点。启动 VMD，在 TkConsole 下输入：

```
mol load pdb membrane.pdb psf membrane.psf
```

可以通过命令行交互模式载入分子，再输入：

```
set mem [ atomselect top all ]
measure center $mem
```

可以得到体系中心的坐标，此例中为 (-22.67，-22.43，-0.76)。然后运行：

```
$mem moveby { 22.67 22.43 0.76 }
$mem writepdb membrane.pdb
```

得到坐标调整的膜坐标文件

载入第 5 小节中生成的 protein.pdb/psf 文件，检查分子结构是否正确。

```
mol load psf protein.psf pdb protein.pdb
```

管子的轴向接近平行于 y 方向，要把它调整到 z 方向去。NAMD 处理半各向同性体系（semi-isotropic）时，默认 z 方向是垂直于各向同性的方向。

```
set sel [ atomselect top all ]
$sel move [ trans center {0 0 0} axis x 95 ]
$sel move [ trans center {0 0 0} axis y 10 ]
$sel writepdb protein _aligned.pdb
```

此时管轴已非常接近于平行 z 轴，若不满足可以通过精心的手调使之平行，或通过 tcl 编程得到倾斜的方向和角度，然后再调整。

move 和 moveby 是 2 个移动体系坐标的命令。move 后面接一个 4×4 的矩阵，可以同时完成平移和旋转操作，但直接书写这个矩阵较为复杂，可以用 trans 生成。moveby 后面接一个方向矢量，完成平移操作。

8．把 GA 嵌入膜中

此处可以使用 Ilya Balabin 提供的 combine.tcl 脚本实现。可以到 http://www.ks.uiuc.edu/Research/vmd/plugins/membrane/下载这个脚本文件。使用此脚本时，TkConsole 当前目录为上述步骤制作的 protein_aligned.pdb、protein.psf、membrane.pdb/psf 文件以及 top_all27_prot_lipid.inp 文件。在控制台输入：

vmd –dispdev text -e combine.tcl

即可把 GA 嵌入膜中，且删除同 GA 重叠的脂和水分子，生成 protein_mem.pdb/psf 文件。

9．补充水

因为自动生成的膜包含的水太少，没有达到使膜充分水化的要求，需要额外在膜两侧加些水。VMD 中有个 plugin 名叫 solvate，可以以多种方式对体系加水。对该体系，在两侧加水的操作为：

solvate protein_mem.psf protein_mem.pdb -minmax {{-25 -25 27} {25 25 40}} -o protein_mem_s1

solvate protein_mem_s1.psf protein_mem_s1.pdb -minmax {{-25 -25 -40} {25 25 -27}} -s WT2 -o protein_mem_s

至此建模操作完成，整个体系约含有 15000 个原子、46 个脂分子。下面即可使用 NAMD 进行分子动力学计算。

solvate plugin 是 VMD 中一个重要的工具，它的工作方式是把分子放入一个足够大的经过充分预平衡的水盒子中，然后删除不符合条件的水分子。

使用方法：

solvat <psffile> <pdbfile> <option1?> <option2?>...

有多种添加水分子的方法可以选择。

① 指定水盒子的位置，如：

solvate foo.psf foo.pdb -minmax {{-25 -25 -40} {25 25 40}}

可以把分子放入边长为 50 Å ×50 Å ×80 Å、中心在原点的盒子中。

② 指定分子 6 个侧面的(-*x*,-*y*,-*z*,+*x*,+*y*,+*z*)，如：

solvate foo.psf foo.pdb -z 10 +z 10

可以在 *z* 方向的两侧各补加 10 Å 厚度的水层，起始位置为此方向的原子坐标的最小/大值。

如果这 6 个方向数值相同，可以输入：

solvat foo.psf foo.pdb -t 10

在 6 个方向都补加 10 Å 水层。

还有其他选项可供选择：

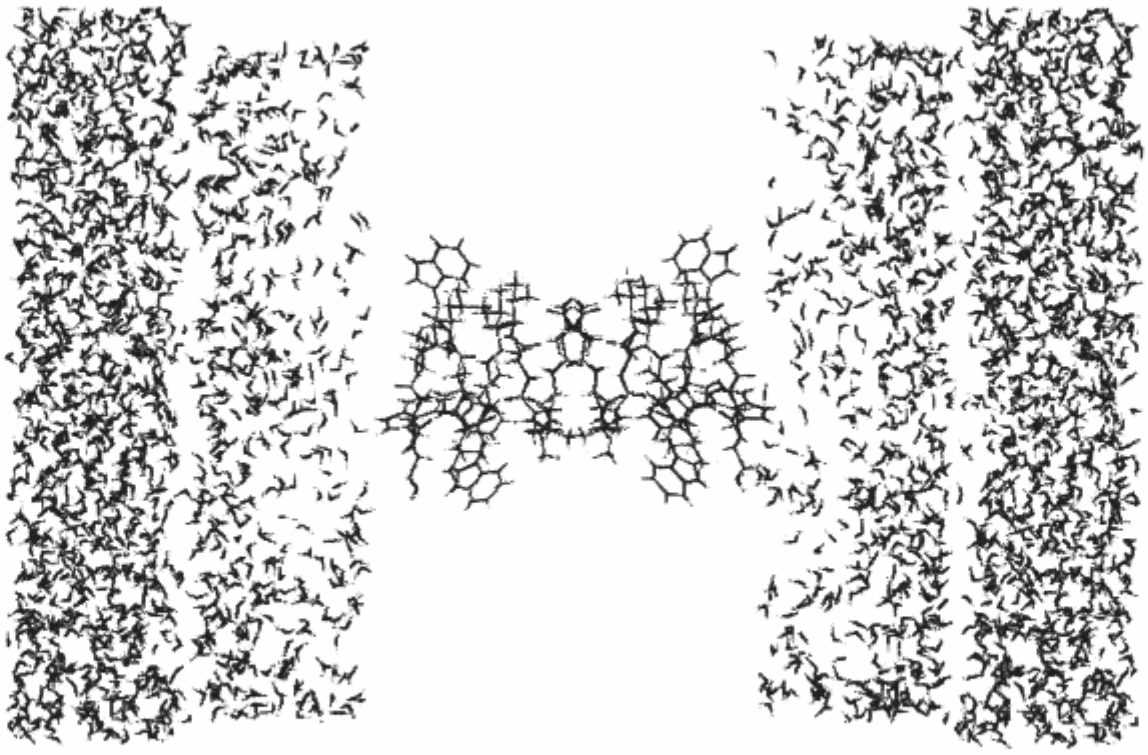
-b（水分子同溶质分子的最短距离）

-s[指定加水的 segment 的名称（默认为 WT1）]

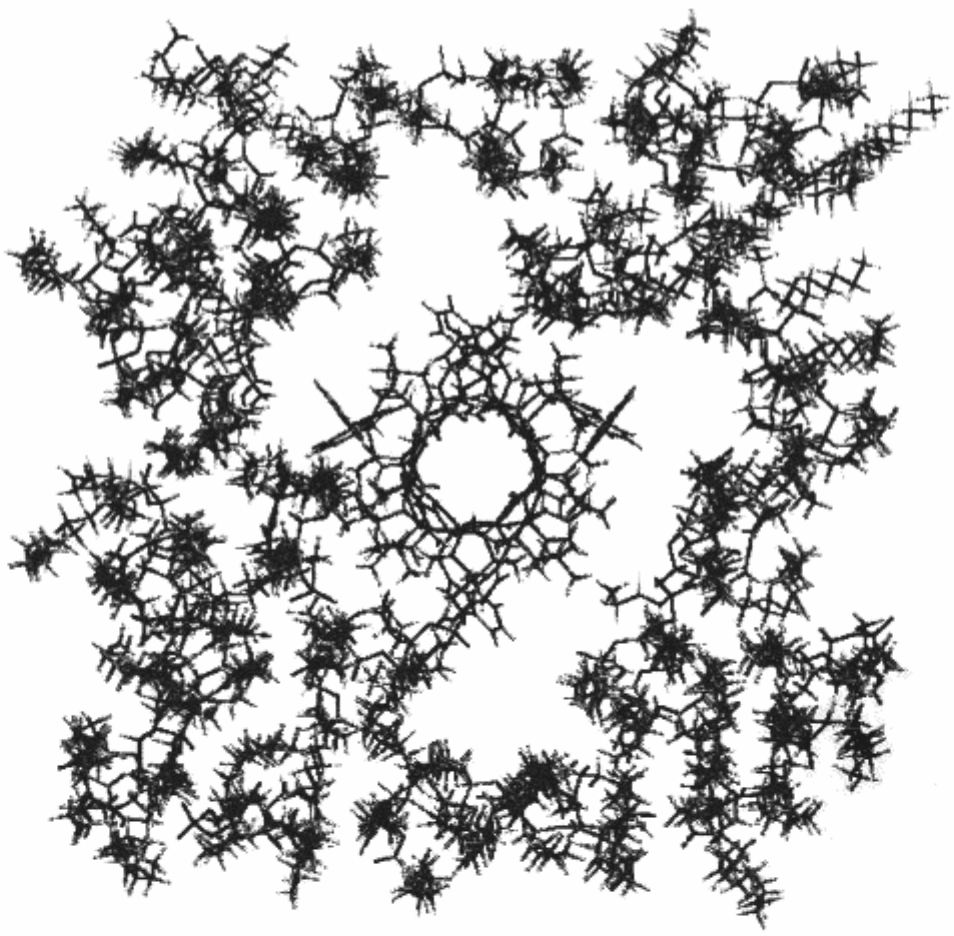
-o（指定生成文件的名称）

如果不指定 psf/pdb 文件，可以生成一个纯的水盒子，此时只能用-minmax 选项。

模型建立后，未进行结构优化时的构象见图 7-20。



（a）侧视图，为使视图清晰，脂分子没有在图中画出



（b）俯视图，水分子没有画出

图 7-20　模型建立后，未进行结构优化时的构象

四、分子动力学模拟

现在建立的模型远远不是平衡态，而只是手工建造的一个体系，有很多不合理的地方，如蛋白质和脂分子之间有很大的空隙。要得到真实的体系，必须进行复杂的预处理过程。因蛋白质的结构取自 XRD 或 NMR 实测值，为蛋白质的天然稳定构象。为了能快速而准

确地完成预平衡过程，需要首先限制蛋白质骨架进行充分的 MD 过程，使得脂和水分子达到平衡。然后全部放开进行长时间的 MD 过程。

因为现在处理的体系较为简单，且为节省篇幅，只做一次能量最小化、限制 C^{α} 的动力学过程和全部放开的动力学过程。实际的处理远远较此复杂。

1．预平衡 1：能量最小化和限制 C^{α} 的结构优化过程

（1）限制 C^{α} 的位置　NAMD 固定/限制某些原子的动力学模拟过程，可以通过标记 pdb 文件中的 occupancy 或 β 列指定所要固定/限制的原子而实现。下列命令可以给 pdb 文件中 C^{α} 的 β 值设为 1，其他设为 0。

```
mol load pdb protein_mem_s.pdb
set all [ atomselect top all ]
set ca [ atomselect top "protein and name CA" ]
$all set beta 0
$ca set beta 1
$all writepdb constr.pdb
```

执行结束后得到文件 constr.pdb，文件内容除 β 列外，其他均同于 protein_mem_s.pdb，用以指定所要限制的原子。

（2）配置文件详解　NAMD 运行所用的绝大部分参数写在一个文件中，这是一个纯文本文件且不依赖于后缀名（后缀名一般为 namd 或 conf），沿用了 Unix 系统下的使用风格。本文采用后缀名 conf，并称之为配置文件或 conf 文件。conf 可以识别 tcl 命令，例如定义变量，判断、循环流程控制等。

能量最小化和限制 C^{α} 的动力学过程可以写在同一个 conf 文件中，文件名 eq1.conf。

主要内容有：

```
structure          protein_mem_s.psf
coordinates        protein_mem_s.pdb
```

指定所要模拟的体系的结构文件和坐标文件。

```
paraTypeCHARMm     on
parameters         par_all27_prot_lipid.inp
parameters         par_extraterms.inp
```

指出所用力场类型为 CHARMm 力场，并指定力场的参数文件名。力场参数文件含有各种类型原子的兰纳德琼斯(LJ)参数以及不同原子之间的键、键角的力常数和平衡值等参数。

本例中带有非标准残基的力场参数文件可从 http://www.ks.uiuc.edu/Research/namd/tutorial/NCSA2002/hands-on/sample/par_all27_prot_lipid.inp 和 http://www.ks.uiuc.edu/Research/namd/tutorial/NCSA2002/hands-on/sample/par-extraterms.inp 处获得。

```
set temperature    310
```

这一句定义了一个变量 temperature 并赋值 310，在下面使用这个数值时候可以引用该变量。

tcl 变量通过开头加$访问。

```
temperature        $temperature
```

设定初始化温度，用以给系统内各个原子按照 Maxwell 分布律赋速度初值。

本例中，初始化温度取自变量 temperature，值为 310，NAMD 的默认温度单位为 K。

cellBasisVector1	50.	0.	0.
cellBasisVector2	0.	50.	0.
cellBasisVector3	0.	0.	80.
cellOrigin	0.	0.	0.
wrapAll	on		

在 NAMD 和 VMD 中，默认的长度单位是 Å。NAMD 不仅能处理正交盒子，还能处理任何夹角的平行六面体盒子，cellBasisVector 是盒子三条边的方向矢量（基矢）。此外，NAMD 还可以处理球形边界和圆柱形边界。

PME	yes
PMEGridSizeX	64
PMEGridSizeY	64
PMEGridSizeZ	100

NAMD 使用 particle mesh ewald（PME）方法计算静电力，此方法需要手工设定划分各自的数目，PMEGridSize 选择略大于相应盒子边长的，且能分解为只包含 2、3、5 约数的正整数。

exclude	scaled1-4
cutoff	10.
switching	on
switchdist	8.
pairlistdist	12.

指定范德华力的截断半径 cutoff、开关函数起始位置 switchdist 和 verlet 列表（第二截断半径）pairlistdist 的数值等。其中 pairlistdict 的数值直接影响计算速度。

timestep	1.
nonbondedFreq	2
fullElectFrequency	4
stepspercycle	20

设定时间步长，单位 fs（10^{-15} s）。

分子动力学模拟的瓶颈是范德华力和静电力的计算，这一部分占总模拟时间的 95%以上。NAMD 采用多时间步长的方式计算这些相互作用，使得计算耗时大大减少。在这个例子中，范德华力每 2 步计算 1 次，静电力每 4 步计算一次，每 20 步更新一次 verlet 列表。

langevin	on
langevinDamping	5
langevinTemp	$ temperature
langevin	控温的参数，使体系温度恒定在 310K（37 ℃）附近。
langevinPiston	on
langevinPistonTarget	1.0
langevinPistonPeriod	100.
langevinPistonDecay	50.
langevinPistonTemp	$ temperature

langevin　　控压的参数，使体系压力恒定在 1 bar（10^5Pa）附近。此处控压的主要目的是调节盒子的大小，使不至于出现真空区。

useFlexibleCell　　yes

useConstantRatio　　yes

这是两个用于膜体系模拟的参数。这个体系为半各向同性体系，当盒子的边长变化时，在平行于膜的两个方向（x、y）上等比例变化。

```
if {1} {
constraints          on
consref              constr.pdb
conskfile            constr.pdb
conskcol             B
constraintscaling    10.0
}
```

这个系列的参数用于限制 C^{α} 原子的运动，即给每个 C^{α} 原子施加偏离 constr.pdb 文件中指定位置的谐力，力常数存储在 constr.pdb 文件的 β 列。因其数值为 1，再乘以因子 10，可得力常数的数值为 10，单位 kcal/Å。

if {1} {… …} 为 tcl 条件控制语句。例子中条件关系式为 1，恒为真，其中的所有设置都会被读入。如果改为 0，则恒为假，所有设置均被忽略。可用此对若干行语句进行开关设置，而不需要逐行注释或删除。

```
binaryoutput         no
outputName           eq1_o
dcdfreq              1000
xstFreq              1000
```

NAMD 有多个输出文件，如坐标（*.coor）、轨迹（*.dcd）、速度（*.vel）、盒子的尺寸（*.xsc/xst）等，这些文件在例子中的名字为 eq1_o。每 1000 步（1ps）存储 1 次轨迹和盒子尺寸。

binaryoutput 选项设定使用 ascii 编码存储*.coor 和*.vel 文件，这类文件采用 pdb 文件格式。

```
restartname          eq1_r
restartfreq          1000
```

在运行中保存一些中间文件，当计算因为某些原因中断（如停电）后，可以根据这些文件继续进行计算。这些文件中包含坐标，速度和盒子尺寸等信息。例子中每 1000 步存储 1 次。

outputEnergies　　1000

NAMD 运行时会输出体系的很多信息，这是最重要的一条，它输出能量、温度、压力、体积等若干很重要的信息。默认每步输出 1 次。现在为 1000 步输出 1 次。

```
IMDon                on
IMDport              12345
IMDfreq              100
```

实现交互分子动力学（interactive molecular dynamic, IMD）的设置，将在下面介绍。

```
if {1} {
minimize                500
reinitvels              $temperature
}
```

首先进行 500 次能量最小化计算，NAMD 提供多种优化方法，选取默认即可。在进行能量最小化后，体系速度为 0。用 reinitvels 命令给体系各个原子重新赋速度初值。

run　　10000000　　进行 10000000（10 ns）步动力学过程。完成后结束。

以上内容是完整进行一次 NAMD 运行过程所需的配置文件的内容，但远非 NAMD 的全部功能。关于 NAMD 可以使用的命令可以参考 NAMD 的用户指南（NAMD user's guide）。当一次过程运行结束后，可以得到如下几个文件：eq1_o.coor（结束时的坐标）；eq1_o.vel（结束时的速度）；eq1_o.xsc（结束时盒子的尺寸）；eq1_o.dcd（运动轨迹）。

（3）运行方式　NAMD 在单个 cpu 下非常简单，只需要运行 namd2　eq1.conf 即可。

单机多核运行，需要开启 rsh 或 ssh 服务，且当前用户可以无密码登录本机。然后运行：

```
charmrun  namd2  +p<n>  eq1.conf
```

<n>为使用的计算核心的数目。Windows 下使用方法同上，但在运行前需要先运行安装目录下的 charmd。

多节点运行，每个节点皆需开启 ssh 或 rsh 服务，且各个节点之间可以无密码登陆。在运行目录下建立 nodelist 文件，在这个文件中指定所使用的节点。然后运行：

```
charmrun  namd2  +p<n>  eq1.conf
```

当使用 ssh 进行通讯时，只能在 csh 下使用，且要设置环境变量：

```
setenv  CONV_RSH  ssh
```

多节点并行时，namd2 文件在每个节点上目录应相同，最好放在公共 NFS 服务器上。如果并行时报错，<configure file> 用绝对路径试试看。

运行时，输出的屏幕（stdout）上的信息也十分重要，需要把这些信息保存在一个文件中。可以用 linux shell 提供的输出重定向功能：

```
namd2  eq1.conf  >  eq1.log
```

可以把原本输出在屏幕上的信息输出到 eq.log 文件中去，或是用：

```
namd2  eq1.conf  |  tee  eq1.log
```

输出信息可以同时输出到屏幕和指定的文件中，并行计算时类似。

log 文件是极重要的输出文件，它采用精心设计的格式，既容易直接阅读，也容易被多种文本操作工具（如 grep、cut 和 awk 等）处理。其中以“ENERGY:”开头的一行包含了模拟过程中体系的能量、温度、压力、体积等多种信息，通过监视这些量的变化可以判断模拟进行的程度。VMD 中有一个 plugin 叫 namdplot，可以很方便地根据这个文件做出某个物理量随时间的变化趋势。打开 extensions | analysis | NAMD plot，以图形界面方式调用这个 plugin。并不是每个 plugin 都有图形界面。在 TkConsole 窗口输入：

```
namdplot  TEMP  eq1.log
```

可以以命令方式调用这个 plugin，并绘出图形。

（4）Interactive MD 的实现　通过监视 log 等文件，可以实时了解体系的状态。而对于想要“看到”体系结构的实时变化，则可以通过 VMD 实现这个功能。连接 NAMD 同 VMD

的协议就是 IMD。IMD 最重要的功能是可以手工在模拟过程中给体系的某一部分施加力，但这种模拟技术需要较深的理论功底才能得到有意义的结论，在这里不介绍。两个软件连用的实现方式是 NAMD 运行时开启一个网络端口做服务器端，等待 VMD 的连接，VMD 作为客户端接入。

在 NAMD 中，与 IMD 有关的主要参数有 3 个:

① IMDon　　<on or off>　　;# 是否开启，默认 off。

② IMDport　　<正整数>　　;# NAMD 开启的网络端口，最好在 1024~65535，并且不要与其他常用软件的端口号相同。在 CMD 窗口使用 netstat -an 命令可以查看已经被其他软件占据的端口号。

③ IMDfreq　　<正整数>　　;# 每隔多少步在 VMD 上更新图像。

在 VMD 端:

① 载入 protein_mem_s.psf 文件，并确保其为 top 分子。

② 打开菜单 Extensions | Simulation | IMD connect，弹出一个新窗口。

③ 在 hostnamd 中填入 NAMD 计算主机（master）的 ip，如果 NAMD 和 VMD 在同一机器上面，这里可以填入 127.0.0.1，在 port 中填入 IMDport 的数值。

④ 点 connect，即可实现 IMD 连接。

若以命令方式调用，可以在 TkConsole 中输入:

```
imd  connect  servername/ip  port
```

2. 预平衡 2：全放开的结构优化过程

接下来放开对蛋白质的 C_α 的限制，进行长时间的动力学模拟过程，使体系充分达到平衡。总的模拟时间为 10 ns。在前一过程的基础上进行下一步的模拟过程，配置文件大部分内容同上，只需要在此基础上修改几个地方即可。

① 将 structure 后的文件名改为 eq1_o.coor;

② 盒子信息按照 eq1_o.xsc 中写入;

③ outputName 和 restartName 分别改为 eq2_o 和 eq2_r;

④ 并把限制 C_α 和能量最小化的 2 个 if 中条件关系式的 1 改为 0。

完成新过程的配置文件的修改，命名为 eq2.conf，然后用前述方法运行即可。在配置文件中，指定这些文件的时候可以使用绝对路径或相对路径引用其他目录下的文件，而不要求一定在配置文件所在的目录下。

因为一个课题需要若干过程才能完成，为使工作有条理、减少出错，可以把每步模拟过程放到一个独立的目录中，例如为建模和 2 步预平衡分别建立目录 build /、eq1/ 和 eq2/，另外一些公共文件，如力场文件、psf 文件，可以单独存放在某个目录中。

3. 平衡分子动力学过程的模拟

当确信一个体系达到平衡态之后，就可以进行正式的分子动力学模拟过程了。为在一定计算资源和时间下，尽可能增长实际的模拟时间，可以采用 2 fs 的时间步长。但此时需要限制 H 原子的伸缩震动。

在配置文件中添加:　　rigidBonds　　All

这个选项会将各个 H 原子与重原子的键长限制在某个数值上。这个数值为力场中规定的平衡键长，而不是初始结构的键长。

然后修改:　　timestep　　2.0

即可设置模拟的时间步长为 2fs，同 1fs 相比可以在相同的计算时间内进行 2 倍实际时间的动力学计算，提高了计算效率。

修改 outputName 和 restartName，分别改为 md_o 和 md_r。

此时进行的模拟为平衡分子动力学过程，使用多的步数只是增加取样量，减少统计误差。因这个例子只是在演示 NAMD 处理分子动力学计算的方法，此处总模拟时间为 20 ns。实际使用时可以选取更长的模拟时间。可以将这个脚本命名为 md.conf，这个任务在一个名为 md 的文件夹中进行。

五、可视化与数据处理

VMD 既是一款分子视图软件，能在图形界面下完成视图、动画以及若干数据处理功能，又内嵌有 tcl 脚本语言，使得它可以以流程控制的方式完成更高级的任务。下面简单介绍几种 VMD 处理数据的方法。

1．图像显示与 render

VMD 能以多种风格进行分子视图处理，从而制作出精致的图像。在本例中较好的操作步骤为：

① 载入 eq2_o.coor 文件，注意此时 file type 选 pdb 而不是使用默认选项；

② 点击菜单 Display | Orthographic，这时视图方式为正交而不是透视；

③ 点击菜单 Display | Axes | Off，隐藏坐标轴；

④ 打开菜单 Graphics | Colors | display | background | 8 white，设置背景为白色；

⑤ 打开菜单 Graphics | Representations，可以设定多个 representation，每个可以指定不同的显示方式，着色方式和需要显示的特定原子的类型。表 7 -1 列出了一个适合本例的组合，效果如图 7-21 所示。

表 7-1 VMD 制作结构图时，各个 representation 的设置

Style	Color	Selection	备　注
Cartoon	Structure	protein or resname FOR or resname ETA	
Bonds	ResName	lipid and noh	Bond Radius 选 0.1
VDW	Name	lipid and（name N or name P）	脂质中的 N 和 P 原子
VDW	Name	water and z > –10 and z < 10	管中的水分子
MSMS	ResName	Water	Material 选 Transparent

除了可以通过截图保存图像外，还可以通过 VMD 提供的多种 render 模式制作图片格式的文件。当调整分子在窗口中的位置后，打开菜单"File | Render "，模式选 snapshot，可以保存为 bmp 格式的图片，但分辨率不可调；其他多种模式都可以任意设定分辨率，比如 tachyon。tachyon 模式除生成 tga 格式的图像外，还生成一个 dat 文件。使用其他辅助工具可以将这个 dat 文件转化为图像格式。如果 VMD 采用默认安装，使用 VMD 软件包中的转换工具的运行方式如下：

/usr/local/vmd-1.8.7beta5/lib/tachyon/tachyon_LINUX –format bmp –res 2000 2000 eq2_0.dat

此即可将 dat 文件转化为高分辨（2000 Pixel × 2000 Pixel）的 bmp 格式的图像。

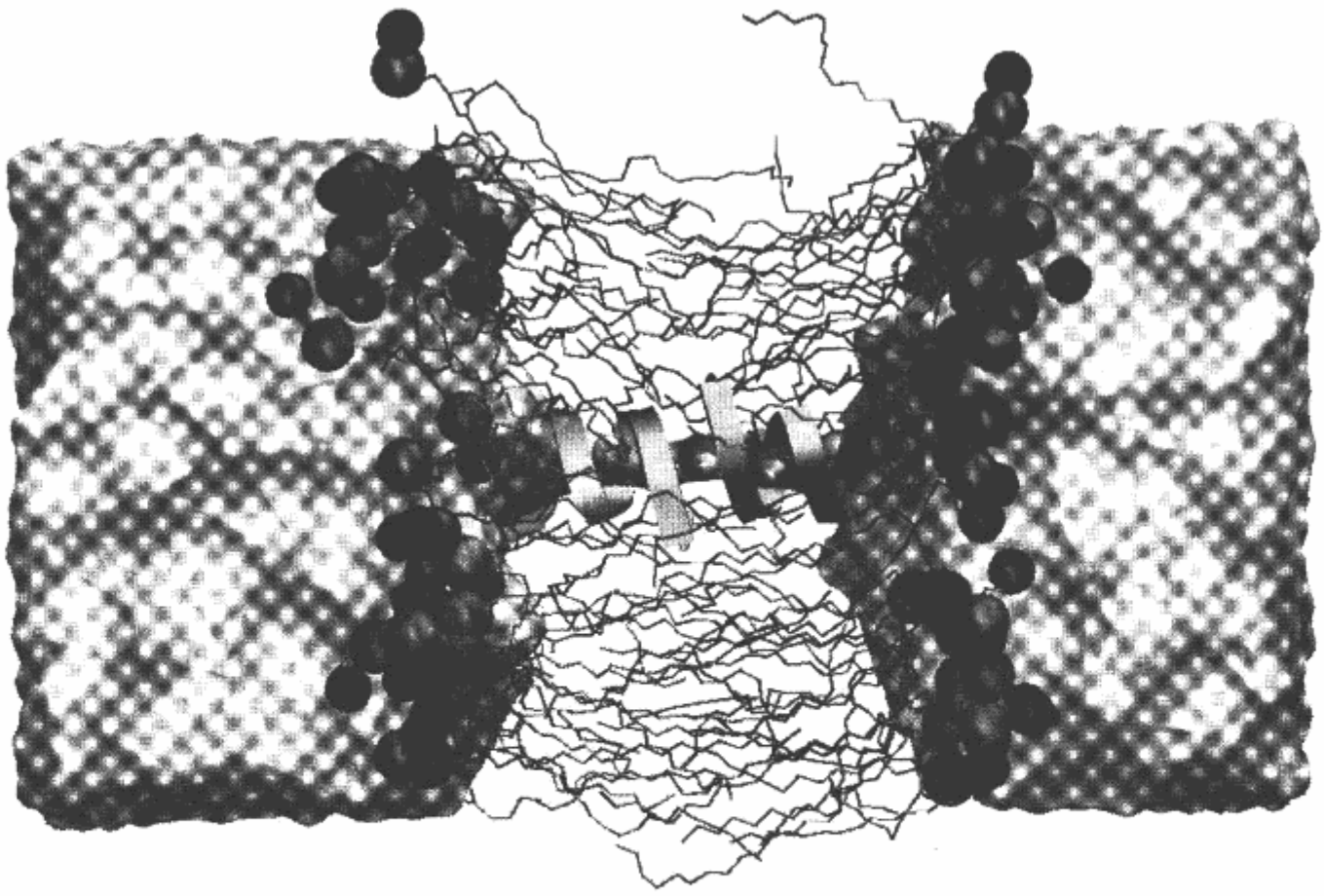


图 7-21　VMD 制作出的体系结构图

2．查看轨迹文件

VMD 除能制作静态的图片外，还可以载入 dcd 文件，以动画方式显示。dcd 文件是 NAMD 产生的轨迹文件，这个文件存储了若干个不同时刻的体系原子坐标（帧，frame）。

首先载入 protein_mem_s.psf 文件，然后载入 eq1_o.dcd 文件，请注意此时“load file for”窗口后面选择这个 psf 文件图 7 -22。在 frames 框中可以设置载入的帧。first 为开始的帧数，0 为开始第一帧；last 为结束的帧数，-1 为文件的最后一帧；stride 可以设定每隔多少帧读取一次。因为要把所有的轨迹数据加载到内存之中，需要花费一定的时间。在前面的模拟中，eq1_o.dcd 的文件大约有 1.7 G，要求系统至少有 1.7 G 的空闲内存，如果内存不够大可以把 stride 设得大一些。在 main 窗口（图 7 -23）下面有 2 行控制动画的工具，功能同视频播放软件类似。

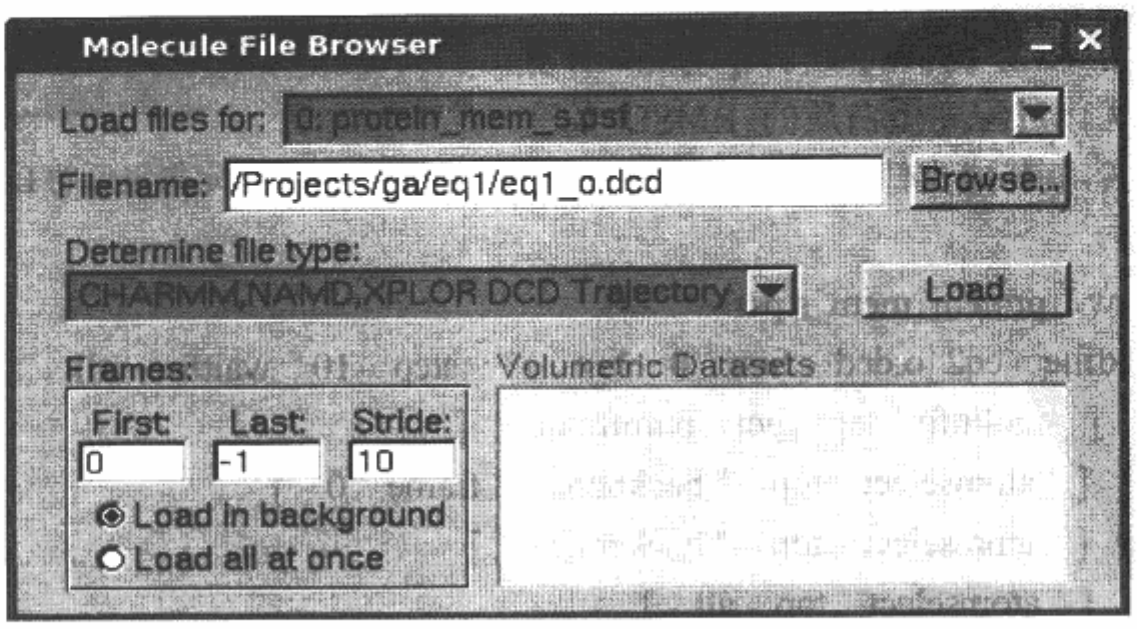


图 7-22　载入轨迹文件的界面

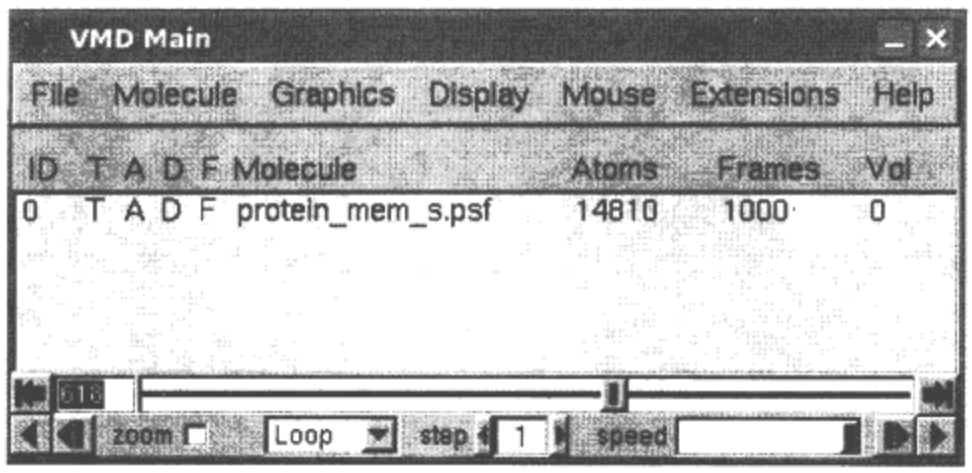


图 7-23　载入轨迹文件后的 main 窗口。

3．使用 Labels 计算键长

Labels 是 VMD 在图形界面使用方式中一个数据分析工具，下面介绍使用 Labels 统计两个原子间距的方法。在两条链的尾段有多条氢键把两个分子组装成二聚体，这里考察每条链的第一个残基（VAL）上的氨基氢同另一条链的第五个残基（ALA）上的羰基氧之间的氢键。

载入 eq2_o.dcd 轨迹文件，在 representations 中设置 selections 为 backbone or resname FOR or resname ETA and noh or name HN，可以较清晰地显示蛋白质的骨架结构。然后点击菜单 Mouse | Labels | Bonds，或在 display 窗口激活后按快捷键“2”，使鼠标变为 Label 模式。在 display 窗口依次选择原子并单击，原子的类型和键的长度会出现在窗口里面。

打开菜单 Graphics | Labels，出现 Labels 窗口（图 7-24）。在下拉菜单中选 Bonds，标签页中选 Graph。选择一个键之后点击 Graph 按钮，可以绘出键长随时间的变化，或按 Save 按钮，可以将数据存储在一个文本文件中。在本例中可以看到这两个氢键键长在1.9 Å左右波动，在模拟过程中没有断裂。

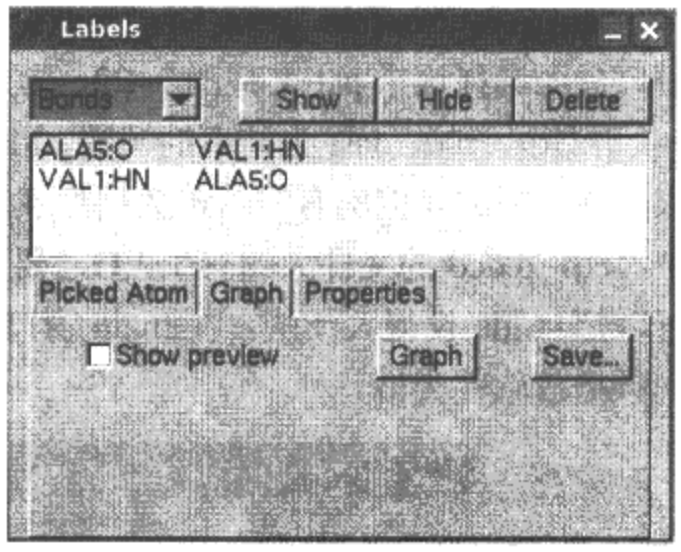


图 7-24　选定氢键后的 Labels 窗口

Labels 中的 Bonds 只是计算两个原子之间的距离，不要求这两个原子真的通过共价键连在一起。像本例中氢键两端的原子就没有键连在一起。

4．使用 tcl 脚本计算蛋白质的 RMSD

通过使用 VMD 提供的 tcl 脚本功能，计算在全部放开优化过程中蛋白质的 RMSD 的变化。将如下内容制作成文件 rsmd.tcl，并放入 eq2 目录中。

```
mol new protein_mem_s.psf
mol addfile eq2_o.dcd first 0 last -1 step 10 waitfor all
set nf [ molinfo top get numframes ]
set ref [ atomselect top "backbone" frame 0 ]
set sel [ atomselect top "backbone" ]
set all [ atomselect top all ]
for { set i 0 } { $i < $nf } { incr i } {
```

```
    $sel  frame  $i
    $all  frame  $i
    $all  move  [  measure  fit  $sel  $ref  ]
    puts  "$i:\t[  measure  rmsd  $sel  $ref  ]"
}
quit
```

使用 VMD 脚本模式运行该文件，将输出重定向到文件中

```
vmd  -dispdev  text  -e  rmsd.tcl  |  tee  rmsd.dat
```

运行结束后就可以使用文本编辑器打开 rmsd.dat 文件查看就可以了。

脚本解析：

```
mol  addfile  eq2_o.dcd  first  0  last  -1  step  10  waitfor  all
```

载入分子。step 10 指每 10 帧载入一次，同载入 dcd 窗口帧的 stride 相同

```
waitfor  all
```

指在轨迹全部加载完毕之后再进行下一步操作

```
set  nf  [  molinfo  top  get  numframes  ]
```

统计轨迹的帧数，并赋给变量 nf

```
set  ref  [  atomselect  top  "backbone"  frame  0  ]
```

定义变量 ref，变量内容为蛋白质骨架原子及其在第一帧时的坐标。vmd 计数从 0 开始

```
set  sel  [  atomselect  top  "backbone"  ]
```

定义变量 sel，变量内容为蛋白质骨架原子，包含各个帧的坐标数据

```
for  {  set  i  0  }  {  $i  <  $nf  }  {  incr  i  }  {  …  }
```

一个遍历从 0 到 numFrame 的循环，循环变量为 i

```
$ref  frame  $i
```

两行

指定当前坐标数据为为第 i 帧构型

```
$all  move  [  measure  fit  $sel  $ref  ]
```

使当前结构同第一帧结构最大限度重合

```
puts  "$i:\t[  measure  rmsd  $sel  $ref  ]"
```

计算 RMSD，并且将当前帧数和 RMSD 输出到屏幕上，每组数据一行，“\t”表示输出一个制表符。

5．使用 bigdcd 计算膜厚度

膜厚度有不同的定义方法，这里采用一个简单的，即计算两侧 P 原子距离的方法。本例中蛋白质结构简单取自 NMR 实验值，且分子较小、结构简单，优化时候很快达到平衡。相对于蛋白质，膜的平衡是较难达到的，可以通过监视膜厚度的方法判断优化进行的程度。对 POPE 膜，通过计算 P 原子距离得到的厚度在 37~38 Å。

对于体积太大不能加载到内存中的脚本，可以使用 catdcd 编辑轨迹文件使体积变小后处理，也可以使用 bigdcd。bigdcd 使用 tcl 语言编写，它提供了一个处理大轨迹的用户接口，其工作原理是每次处理完一个 frame 后即把它从内存中删除。从 http://www.ks.uiuc.edu//Research/vmd/script_library/下载 bigdcd.tcl 文件，编辑如下脚本，命名为 thickness.tcl。

```
source  bigdcd.tcl
set  outfile  [  open  thickness.dat  w  ]
```

```
proc myth { frame } {
  global mol outfile
  set p1 [ atomselect top "name P and z > 0" ]
  set p2 [ atomselect top "name P and z < 0" ]
  set z1 [ lindex [ measure center $p1 ] 2 ]
  set z2 [ lindex [ measure center $p2 ] 2 ]
  set th [ expr $z1 - $z2 ]
  puts $outfile "$frame\t$th"
}
set mol [ mol new protein_mem_s.psf waitfor all ]
bigdcd myth eq1/eq1_o.dcd eq2/eq2_o.dcd
bigdcd_wait
close $outfile
quit
```

运行结束后，会自动生成一个新的文本文件 thickness.dat，可以使用任何一款文本编辑器查看或是用数据处理软件如 xmgrace 、excel 等编辑作图。图 7-25 为本例中的结果。

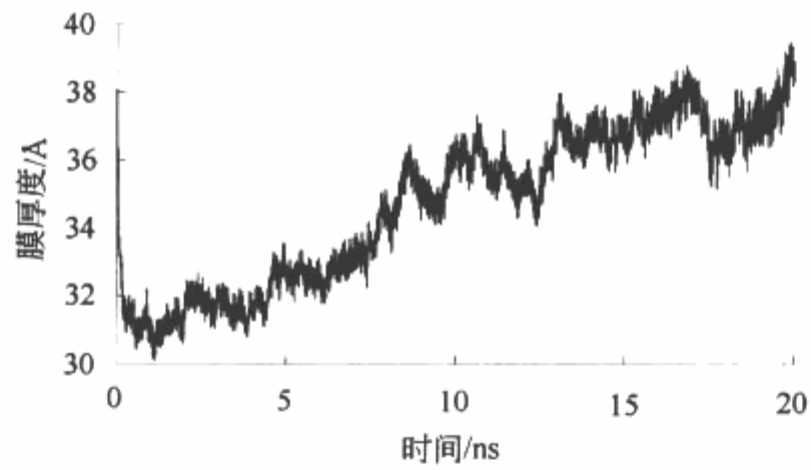


图 7-25 膜厚度在 2 次预平衡过程中随时间的变化趋势

在最先部分膜厚度急剧下降后缓慢上升，10ns 时候（预平衡 1 结束时）达到期望值 37 Å。后 10ns（预平衡 2）过程在期望值附近波动。可以认为经过两次预平衡，体系充分达到了平衡状态

脚本解析:

```
source bigdcd.tcl
```

载入 tcl 脚本，使其中定义的函数接口生效

```
set outfile [open thickness.dat w]
```

定义一个文件变量，变量名 outfile

```
proc myth { frame } { ... }
```

定义一个函数，用于处理单帧数据。函数的变量为 frame，由 bigdcd 传递，是当前帧数

```
global mol outfile
```

设置 mol 和 outfile 为全局变量，定义它的位置在这个函数 myth{}之外

```
set z1 [ lindex [ measure center $p1 ] 2 ]
```

两行

计算某侧 P 原子的坐标中心，并将 z 坐标赋给 z1 和 z2

```
set th [ expr $z1 - $z2 ]
```

```
    计算 z1 和 z2 之差，赋给变量 th，th 就是要统计的膜的厚度
    puts  $outfile  "$frame\t$th"
    输出当前帧数和膜厚度到 thickness.dat 文件中
    bigdcd  myth  eq1/eq1_o.dcd  eq2/eq2_o.dcd
    bigdcd_wait
    使用 bigdcd 的语法，bigdcd 后接上面定义的函数名，再接 dcd 轨迹文件，可以只有一个，也可以同时跟多个，顺序由前到后
    close  $outfile
    关闭文件
```

第四节　利用 GROMACS 进行分子动力学模拟操作

一、GROMACS 介绍

GROMACS 全称为 Groningen Machine for Chemical Simulations，是一款最初由荷兰罗宁根大学（Groningen University）开发的开源分子动力学模拟软件包，现在主要由瑞典乌普萨拉大学（Uppsala University）和斯德哥尔摩大学（Stockholm University）以及德国马普学会高分子研究所（The Max Planck Institute for Polymer Research）共同维护管理。

GROMACS 设计的初衷主要是为了用于生物大分子如蛋白质、DNA、RNA 等物质的计算、分析。后来由于其在模拟大分子系统的牛顿运动方面具有极大的速度优势，很多研究机构也用其来计算、研究非生物大分子体系，如有机或无机高分子材料。GROMACS 主要是用来研究分子体系的动力学性质，它可以用分子动力学、随机动力学或者路径积分方法模拟溶液中分子热动力的性质，进行分子能量的优化、构象分析等。GROMACS 支持很多算法，而且与其他动力学模拟软件相比还具有其独特的优势。

① 为了提高计算能力，GROMACS 的设计者们对算法进行了大量的优化，使得它成为目前主流的分子动力学模拟软件之一。但是要特别强调的是，GROMACS 大大地提高了计算速度，同时并不以牺牲计算精度为代价。

② GROMACS 的用户界面设计非常人性化，而且操作简单、功能丰富，对于初学者来说易于上手。同时拓扑文件和参数文件都以文档的形式给出，便于编辑。在程序运行过程中，并不用输入脚本注释语言。而且运行的过程是分步进行的，可以随时检查模拟的正确性和可行性，有效地减少了时间的浪费。

③ GROMACS 的轨迹存储采用了损耗压缩的方法，这一点相比于同类的动力学模拟软件有很大的区别，可以为使用者提供了一个压缩性很强的轨迹数据文件，并且压缩的精度可以由用户来选择。

④ GROMACS 源码公开，是一个很好的开源动力学模拟软件。

下面就以 Unix/Linux 版的 GROMACS 软件为例，讲解其安装、使用以及分析等具体操作。

二、GROMACS 的下载及安装

本实验以 Redhat Advanced Server 4 Update 7 操作系统为例来讲述 GROMACS-3.3.3 的安装及运行过程。所需文件：fftw-3.0.2.tar.gz 和 gromacs-3.3.3.tar.gz。

系统要求：C 编译器和 Fortran 编译器。由于 GROMACS 在编写的时候用到了 fftw 函数

库，因此需要首先下载安装 fftw。

① 到 http://www.fftw.org/下载 FFTW 的源代码。

② 执行以下命令解压缩并进行编译安装：

```
$ mkdir $HOME/gromacs
$ mkdir $HOME/gromacs/fftw-3.0.2
$ tar -xzvf fftw-3.0.2.tar.gz
$ cd fftw-3.0.2
$./configure --enable-float --enable-threads --prefix=$HOME/gromacs/fftw-3.0.2
$ make
$ make install
```

然后把以下两行写在当前用户目录下的 bashrc 文件中，保存退出并运行“bash”。

```
$ export CPPFLAGS=-I$HOME/gromacs/fftw-3.0.2/include
$ export LDLAGS=-L$HOME/gromacs/fftw-3.0.2/lib
```

fftw 安装完毕，紧接着安装 GROMACS。

③ 到 http://www.gromacs.org/下载 GROMACS 的源代码，这里以 GROMACS-3.3.3 为例。

④ 执行以下命令解压缩进行编译安装：

```
$ mkdir -p $HOME/gromacs/gromacs-3.3.3
$ tar -xzvf gromacs-3.3.3.tar.gz
$ cd gromacs-3.3.3
$ ./configure --prefix=$HOME/gromacs/gromacs-3.3.3 --with-fft=fftw3
$ make
$ make install
```

到现在就已经安装成功了，GROMACS 被安装在指定的$HOME/gromacs/grom-acs-3.3.3 目录中。然后把下面这一行写在当前用户目录下的 bashrc 文件中，保存退出并运行“bash”。

```
$ export PATH=$PATH: $HOME/gromacs/gromacs-3.3.3/bin
```

⑤ 参考 online 的 tutorial 运行$HOME/gromacs/share/tutor 目录下的教程体验，运行 luck，会出现名言名句，说明安装成功。

三、GROMACS 的基本操作

在介绍 GROMACS 的基本操作之前，有必要首先对 GROMACS 中的一些常用文件类型进行介绍。①molecular topology files，这是 GROMACS 的分子拓扑文件（top file），用 GROMACS 自带的子程序“pdb2gmx”可以自动将 PDB 格式文件转化生成 top 文件。②molecular structure files，这种结构文件包括 pdb 和 gro 两种，用“pdb2gmx”将 pdb 文件转化为 top 文件的同时也可以将 pdb 转化为 gro 文件。gro 和 pdb 都是结构文件，它们的不同在于格式，gro 文件除了保存原子的坐标外，还可以保存动力学模拟中的盒子尺寸、速度等信息。关于模拟中的盒子信息，主要是通过 GROMACS 的另外一个子程序“genbox”来加入的。③molecular dynamics parameter files (mdp)，mdp 文件其实是一个参数文件，包括了要做分子动力学模拟所需的参数例如时间、步数、温度、压力、速度等。④index files(ndx)，GROMACS 的索引文件，即 index 文件，由 make_ndx 子程序生成，文件后缀为 ndx。索引文件是 GROMACS

非常重要的文件，使用它可以使动力学模拟变得随心所欲。举一个简单的例子，比如想做QM/MM分子动力学，即对于结构中比较关键的部分，采用量子力学方法，而对于其他部分，用分子力学的方法处理。于是就面临一个对模拟系统进行分割定义的问题，这时索引文件就要发挥作用了。大体的思路是，在索引文件中，另外定义一个单独的组，该组中包括需要用量子力学处理的所有原子。在mdp文件中，对这个组定义用量子力学来模拟。具体的信息可以参考GROMACS的使用说明。⑤run input file (tpr)，用GROMACS的子程序grompp将上面提到的四种文件合并生成tpr文件，该文件包括了分子动力学模拟所需要的所有信息。⑥trajactory files (trr)，这就是动力学模拟时所生成的轨迹文件，它记录了模拟过程中一定时间间隔的瞬时结构信息，它可以用GROMACS中的子程序“ngmx”来读取，当然也可以用另外一个叫做VMD的软件读取。读取之后，可以像看电影一样观察结构随时间的变化状况。

用GROMACS进行分子动力学模拟大致可以分成以下5个操作步骤：初始文件的准备；初始结构的优化；分子动力学模拟；结果分析。

接下来以GID1蛋白（pdb编号为：2ZSH）为例对GROMACS的操作步骤进行详细的介绍，把GID1的pdb文件名命名为complex.pdb。

1．初始文件的准备

① 将pdb文件转换生成GROMACS的结构文件（*.gro）和拓扑文件（*.top），执行以下命令：

```
pdb2gmx –ff oplsaa –f complex.pdb –o complex.gro –p complex.top > pdb2gmx.log
```

在进行分子动力学模拟之前，必须建立分子的拓扑文件。运行pdb2gmx后会要求用户选择合适的分子力场（图7-26），关于力场的选择可以参考GROMACS手册中相关的章节。这一步将会由原始的pdb文件得到结构文件以及相应的拓扑文件，二者都可以使用文本编辑器进行查看。由于从www.pdb.org上下载得到的PDB文件都不含有氢原子，pdb2gmx可以添加所有缺失的氢原子。在pdb2gmx输出的GROMACS结构文件中，包含了蛋白质结构的每一个原子，并定义了分子结构的尺寸大小，文件后缀名为gro。pdb2gmx.log是pdb2gmx命令输出文件。此外，还有一个posre.itp文件，被称为“索引文件”。对于pdb2gmx程序的其他用法请参阅GROMACS的参考手册或运行：

```
pdb2gmx -h
```

```
Opening library file /data2/soft/gromacs-4.0.2/share/gromacs/top/FF.dat

Select the Force Field:
 0: GROMOS96 43a1 force field
 1: GROMOS96 43a2 force field (improved alkane dihedrals)
 2: GROMOS96 45a3 force field (Schuler JCC 2001 22 1205)
 3: GROMOS96 53a5 force field (JCC 2004 vol 25 pag 1656)
 4: GROMOS96 53a6 force field (JCC 2004 vol 25 pag 1656)
 5: OPLS-AA/L all-atom force field (2001 aminoacid dihedrals)
 6: [DEPRECATED] Gromacs force field (see manual)
 7: [DEPRECATED] Gromacs force field with hydrogens for NMR
 8: Encad all-atom force field, using scaled-down vacuum charges
 9: Encad all-atom force field, using full solvent charges
```

图7-26　运行pdb2gmx命令后的示意图

② 在水箱中对分子进行溶剂化。由于实际状况下所要模拟的蛋白质的行为都是在水溶液中发生的，所以模拟之前，必须给分子添加水环境。GROMACS中这一步通过使用editconf

和 genbox 两个子程序实现。前者用来生成一个包裹整个分子的长方体空箱，后者则会读取结构文件并向其中加入水。Genbox 命令读入 GROMACS 的结构文件，并读入水盒子尺寸的大小，输出文件包含分子文件和水盒子。同时 genbox 更改原来的拓扑文件，使其包含水分子。在使用 genbox 之前，要使用 editconf 命令定义水盒子大小，执行以下命令：

editconf -f complex.gro -o complex.gro -bt octahedron -c -d 0.8 > genbox.log

“-d 0.8”指的是从蛋白质边缘到盒子边缘的距离是 0.8nm，大约是 8Å，一般来说在 0.5～0.8。“-c” 指的是将蛋白质放在盒子的中央。特别要注意的是，editconf 这个子程序也可以用来将 gro 文件转化成 pdb 文件（相当于 amber 中的 ambpdb 命令）：

editconf -f file.gro -o file.pdb

接着运行以下命令向盒子中加入水分子：

genbox -cp complex.gro -cs spc216.gro -p complex.top -o complex_wat.gro >> genbox.log

genbox 还会修改相应的 top 文件，执行：

tail complex.top

即可看到 genbox 加入的溶剂分子的数目。

原则上讲，到此为止已经可以开始进行模拟计算了。但是，在准备模拟初始文件的时候，体系中有很多原子距离太近，局部能量过高。这些相互距离太近的原子大多是由于 genbox 程序产生的，溶剂分子与蛋白分子之间存在不稳定的高能量。如果现在开始模拟计算，而不进行能量最优化，系统将可能很不稳定。能量最优化过程是改变系统中局部高能量的原子的位置，降低这些点的能量。

2．初始结构的能量最优化

在进行能量最优化之前，先用 GROMACS 的预处理程序 grompp 处理所有输入文件。grompp 预处理拓扑文件（top）、坐标文件（gro）和一个参数文件（mdp），然后输出一个二进制拓扑文件（tpr）。这个二进制文件包含所有模拟需要的信息，利用这个文件就可以进行能量最优化和动力学模拟。这里给出了一个进行结构优化的 mdp 参数文件，其中所涉及的相关项的解释可以见附录或者参考 GROMACS 的手册。运行以下命令：

grompp -v -f em.mdp -c complex_wat.gro -p complex.top -o em.tpr

em.mdp 参数文件范例

```
    title = minimization
cpp = /usr/bin/cpp
define = -DFLEX_SPC
constraints = none
integrator = steep
nsteps = 5000
emtol = 10.0
emstep = 0.01
nstlist = 10
ns_type = grid
;Using PME
coulombtype = PME
rlist = 0.9
```

rcoulomb = 0.9

rvdw = 0.9

pme_order = 4

fourierspacing = 0.12

fourier_nx = 0

fourier_ny = 0

fourier_nz = 0

ewald_rtol = 1e–5

optimize_fft = yes

有了二进制拓扑文件，现在可以开始进行能量最优化了。进行能量最优化的程序是mdrun，这也是 GROMACS 的核心程序，所有的模拟都是用 mdrun 程序进行的，就好像GROMACS 的心脏一样，类似于 amber 软件中的 sander 程序。不过在进行优化之前还有一件非常重要的事情，就是要首先向体系中加入抗衡离子来使体系达到中性，因为所要模拟的生物体系中的某些变化一般都是在接近中性的条件下发生的。这里所研究的体系显负电性，因此需要加入钠离子来使其达到电中性。运行以下命令：

genion -s em.tpr -o complex_wat_ion.gro -nname Na -np 2 -p complex.top -g complex_wat_ion.log

完成后，就可以进行能量最优化了。在进行能量最优化过程中，注意查看 mdrun 程序的输出文件。在输出文件当中，Step 是模拟步数，Time 是模拟时间，Energies 是系统的各种能量值（图 7-27）。如果模拟顺利，可以看到系统势能从一个很高的值迅速降低，最后稳定在一个较低的值。运行以下命令：

mdrun -v -s em.tpr -o em.trr -c em.gro -e em.edr -g em.log

```
           Step             Time           Lambda
              4          4.00000          0.00000

   Energies (kJ/mol)
           Bond            Angle    Proper Dih.  Improper Dih.          LJ-14
    6.86825e+03    2.39288e+03    1.00113e+03    3.01632e+02    4.43385e+03
     Coulomb-14        LJ (SR)   Coulomb (SR)   Coul. recip.      Potential
    3.53141e+04    1.54032e+05   -8.33652e+05   -1.33460e+05   -7.62767e+05
Pressure (bar)
    2.48345e+04

           Step             Time           Lambda
              5          5.00000          0.00000

   Energies (kJ/mol)
           Bond            Angle    Proper Dih.  Improper Dih.          LJ-14
    4.95299e+03    2.67851e+03    9.98294e+02    2.33302e+02    4.05909e+03
     Coulomb-14        LJ (SR)   Coulomb (SR)   Coul. recip.      Potential
    3.53769e+04    1.52205e+05   -8.36492e+05   -1.33802e+05   -7.69790e+05
Pressure (bar)
    2.42789e+04
```

图 7-27　体系能量优化的输出文件

优化最后一共可以得到 4 个文件：em.gro、em.trr、em.log 和 em.edr。第一个是优化后得到的结构文件，第二个是优化得到的轨迹文件，第三个是优化的记录文件，最后一个则是系统能量变化。在这里可以运行 editconf 程序将 gro 文件转化为 pdb 文件，并用三维结构可视

化软件，比如 pymol 或 VMD 读入 pdb 文件，来查看优化后分子的构象。运行命令：editconf -f em.gro -o em.pdb

3．水平衡分子动力学模拟

能量最优化之后，首先保持蛋白质不动，对蛋白质周围水环境进行动力学模拟，该过程称为位置限制性分子动力学（position restrained MD），也叫水平衡分子动力学模拟。即保持蛋白质位置不变，对溶剂分子进行平衡计算，这样做的目的是为了填补溶剂空洞，这些空洞产生的原因是由于 genbox 程序加溶剂不均匀造成的。这时就需要对相应的 top 文件进行修改以加入束缚条件。发现在 complex.top 中有这样几行：

```
#ifdef POSRES
#include "posres.itp"
#endif
```

事实上这是一个条件包含语句，仅仅当变量 posres 的值被设置时，才会包含 posres.itp 中的内容，关于 itp 文件其实之前也谈到了它是一个索引文件。这样就不需要修改 top 文件，使用相同的 top 文件也可进行自由状况下的分子动力学模拟。因此，只要在相应的 mdp 文件中对 posres 变量的值进行设置即可实现束缚状态下的动力学优化。

接下来要做的是对输入文件进行预处理得到二进制拓扑文件，这些输入文件就是能量最优化得到的输出文件。然后还有一个参数配置文件和索引文件。默认情况下，程序对模拟系统是分部进行的。利用两个部分进行位置限制性模拟：Protein 和 SOL 部分，分别表示蛋白质和溶剂。在这个过程中，即保持 protein 位置不变。参数文件（mdp）包含了位置限制性模拟的所有参数，包括步长、步数、温度等。同时参数文件告诉 GROMACS 模拟的类型，如能量最优化、位置限制性或者是分子动力学模拟。

pr.mdp 参数文件范例

```
title = pr
cpp = /usr/bin/cpp
define = -DPOSRES
constraints = all-bonds
integrator = md
dt = 0.002  ; ps !
nsteps = 20000   ; total 40 ps.
nstcomm = 1
nstxout = 500 ;collect data every 1 ps
nstvout = 500
nstfout = 0
nstlog = 500
nstenergy = 500
nstlist = 10
ns_type = grid
rlist = 0.9
coulombtype = PME
rcoulomb = 0.9
```

```
rvdw = 0.9
fourierspacing = 0.12
fourier_nx = 0
fourier_ny = 0
fourier_nz = 0
pme_order = 4
ewald_rtol = 1e-5
optimize_fft = yes
; Berendsen temperature coupling is on in two groups
Tcoupl = berendsen
tc-grps = Protein Other
tau_t = 0.1 0.1
ref_t = 300 300
; Energy monitoring
energygrps= Protein Other
; Pressure coupling is on
Pcoupl = berendsen
tau_p = 0.5
compressibility = 4.5e-5
ref_p = 1.0
; Generate velocites is on at 300 K.
gen_vel = yes
gen_temp = 300.0
gen_seed = 173529
```

同之前的优化类似，运行以下命令产生 tpr 文件：

```
grompp -v -f pr.mdp -c em.gro -p complex.top -o pr.tpr
```

然后运行：

```
mdrun -v -s pr.tpr -o pr.trr -c pr.gro -e pr.edr -g pr.log
```

4．无限制的分子动力学模拟

到此为止，分子系统已经可以进行完全无限制的分子动力学模拟了。再一次利用 grompp 程序将输入文件转化成 tpr 预处理文件。

md.mdp 参数文件范例

```
title = md
cpp = /usr/bin/cpp
constraints = all-bonds
integrator = md
dt = 0.002 ; ps !
nsteps = 500000 ; total 1000 ps.
nstcomm = 1
nstxout = 500 ;collect data every 1 ps
```

```
nstvout = 0
nstfout = 0
nstlog = 500
nstenergy = 500
nstlist = 10
ns_type = grid
rlist = 0.9
coulombtype = PME
rcoulomb = 0.9
rvdw = 0.9
fourierspacing = 0.12
fourier_nx = 0
fourier_ny = 0
fourier_nz = 0
pme_order = 4
ewald_rtol = 1e-5
optimize_fft = yes
; Berendsen temperature coupling is on in two groups
Tcoupl = berendsen
tc-grps = Protein Other
tau_t = 0.1 0.1
ref_t = 300 300
; Energy monitoring
Energygrps = Protein Other
; Pressure coupling is on
Pcoupl = berendsen
tau_p = 0.5
compressibility = 4.5e-5
ref_p = 1.0
; Generate velocites is on at 300 K.
gen_vel = no
gen_temp = 300.0
gen_seed = 173529
```

运行以下命令产生 tpr 文件：

```
grompp -v -f md.mdp -c pr.gro -p complex.top -o md.tpr
```

然后运行如下命令，到这里真正的分子动力学模拟就开始了，这个模拟一般需要很长的时间。

```
mdrun -v -s md.tpr -o md.trr -c md.gro -e md.edr -g md.log
```

5．结果分析

动力学最重要的部分不是模拟而是分析，从模拟的轨迹中提取有用的信息，并加以分析

解释现象才是最关键的。在这里主要把 GROMACS 经常用到的几个分析工具介绍给大家。

动力学模拟做完以后，首先可以使用 GROMACS 的 ngmx 程序观看动力学模拟轨迹（图 7-28），轨迹文件包含了分子动力学过程的坐标、原子速度和受力情况等。比如某一个动力学模拟了大概 1ns，就可以用 ngmx 来观察 1ns 的时间段内，体系中各个残基的构象变化情况。只需要运行如下的命令：

```
ngmx -s md.tpr -f md.trr
```

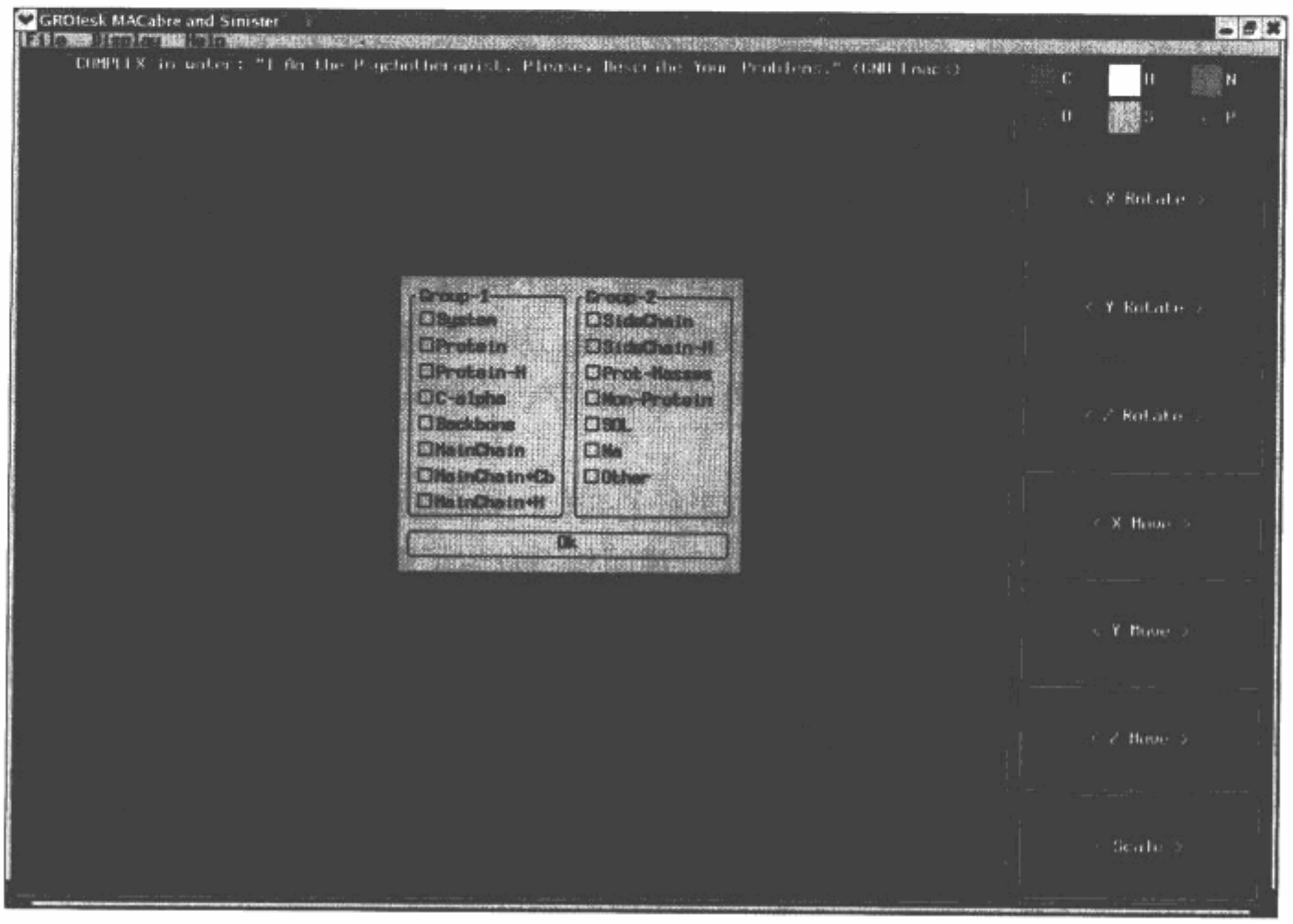


图 7-28　ngmx 的运行界面

其实作为一个开源的分子动力学软件，GROMACS 已经提供了相当多的分析程序，一般来说，动力学模拟完成后要做的第一件事就是看一下该体系是否达到了力平衡。一般用 g_rms 这个子程序来计算蛋白质中各个部分的 RMSD 值（即均方根偏差）随模拟时间的变化情况，只有当体系的 RMSD 趋于稳定时，动力学模拟才达到了平衡。同时可以配合 xmgrace 软件进行作图，这样可以更直观地评价体系的平衡状况。例如利用 g_rms 计算该体系（GID1）中骨架原子的 RMSD 值随时间的变化情况（图 7-29）。

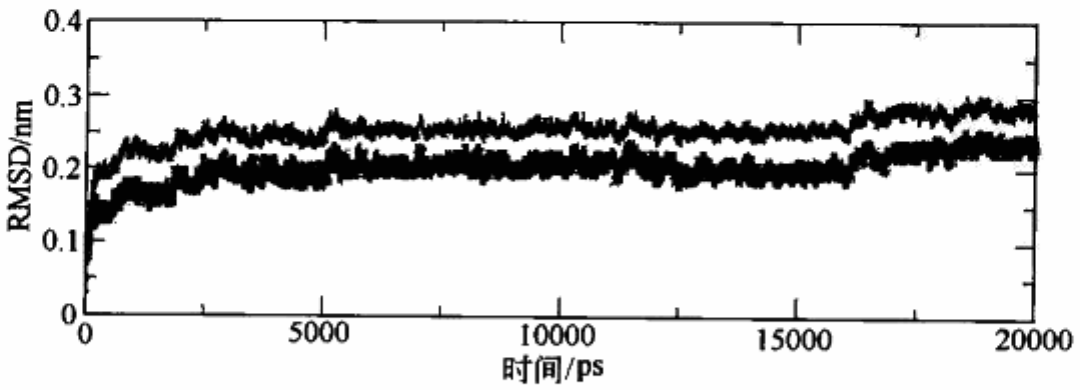


图 7-29　骨架原子的 RMSD 随模拟时间的变化情况

（上面部分是整个蛋白质的骨架，下面部分是蛋白质 N 端区域的骨架）

只分析体系的 RMSD 值往往是不够的，同时还要其他几款子程序来评价体系的平衡状况，可以利用 g_energy 提取体系的势能随模拟时间的变化情况，因为随着模拟时间的增加，体系趋于平衡，往往势能也是趋于平衡的。

也可以利用 g_dist 和 g_bond 评价某些关键原子之间的距离，因为当体系平衡时，这些关键的距离也是趋于恒定的。g_dist 的 index 写法是，在里面加入两个 group，内容分别是这两个原子编号，比如运行 g_dist -f md.xtc -s md.tpr -n index.ndx，之后会提示选择 group，分别选择那两个 group 就行了，结果输出在 dist.xvg。

一般来说，当一个体系的 RMSD 趋于稳定，势能趋于稳定并保持最低，关键距离保持稳定，基本可以断定该体系达到了模拟平衡，可以利用平衡的轨迹进行深层次的分析。以下介绍一些用得比较多的分析工具。

g_rmsf：如果想要知道所模拟的蛋白质的各部分的氨基酸残基在动力学模拟过程中随时间的波动情况，就要用到这个软件了。通常来说，蛋白质保守区的波动性较小，端基和 loop 区的波动性较大，利用该程序，往往可以观察到蛋白质在溶液中的某些特殊变化，有时也会有令人意想不到的发现，同时该程序也可以取动力学一定时间段内的平均结构，是一款很不错的分析工具。

g_gyrate：可以用来测量蛋白质回旋半径的变化，蛋白质的回旋半径反应了蛋白质分子的体积和形状。同一体系的回旋半径变大，说明体系发生了膨胀。

g_sas：可以用来分析溶剂的可及表面，溶剂可及表面是描述蛋白质疏水性的重要指标，氨基酸残基的疏水性是影响蛋白质折叠的重要物理性质。同时体系的溶剂可及表面的变化也会影响溶剂化能。

do_dssp：可以用来分析蛋白质的二级结构变化，往往在研究肽链折叠的分子动力学时要用到这个工具。

g_hbond：用来测量氢键数目、距离和角度随时间的变化。

g_saltbr：用来分析带电残基之间的盐桥的距离随时间的变化。

g_confrms：此程序可以将模拟的最终结构与初始结构进行叠合比较，用来分析动力学后的结构与晶体结构的差异。

虽然 GROMACS 带了很多的分析程序，但是唯一不足的是，当模拟药物与蛋白质之间相互作用时，缺乏两者之间吉布斯结合自由能的精确计算工具，比如像 Amber 程序中所带的 MM_PBSA。如果想要评价 GROMACS 动力学模拟后药物的结合能，就需要自己写相关的脚本程序来实现了，由于涉及比较深的知识，所以这里不做介绍。

本书主要参考文献

[1] 邹承鲁. 第二遗传密码. 长沙：湖南科学技术出版社，1997.

[2] 陈惠鹏. 医药生物工程进展. 北京：人民军医出版社，2004.

[3] 来鲁华等. 蛋白质的结构预测与分子设计. 北京：北京大学出版社，1993.

[4] 王恒樑等译. 蛋白质组学原理. 北京：化学工业出版社，2007.

[5] Eisenhaber F, Press B, Argos P. Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. Crit Rev Biochem Mol Biol, 1995, 30(1)：1-95.

[6] Al-Lazikani B, Jung J, Xiang Z, Honig B. Protein structure prediction. Curr Opin in Chem Biol, 2001, 5：51-56.

[7] Jones D T. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol, 1999, 292: 195-202.

[8] Blundell T L, et al. Nature, 1987：326：347.

[9] Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 2006, 22：195-201.

[10] Luthy R, Bowie J U, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature, 1992, 356：83-85.

[11] Jones D T, Taylor W R, Thornton J M. A new approach to protein fold recognition. Nature, 1992, 358：86-89.

[12] http://www.loria.fr/~ritchied/hex/manual/hex_manual.html.

[13] http://autodock.scripps.edu/faqs-help/tutorial.

[14] Venkatachalam C M, Jiang X, Oldfield T, Waldman M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Mod, 2003, 21：289.

[15] Bhandarkar M, Bhatele A, Bohm E, Brunner R, Buelens F, et al. NAMD User's Guide Version 2.7b1. The Board of Trustees of the University of Illinois, 2009.

[16] Humphrey W, Dalke A, Schulten K. VMD：visual molecular dynamics. J Mol Graphics, 1996, 14：33-38.

[17] http://www.gromacs.org/.

[18] http://www.umdnj.edu/~kerrigje.